

Bond University

DOCTORAL THESIS

Clinical prediction rules for assisting diagnosis

Sanders, Sharon L

Award date:
2016

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.



Clinical prediction rules for assisting diagnosis

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

Sharon Lea Sanders

Centre for Research in Evidence Based Practice

Faculty of Health Sciences & Medicine

Bond University, Australia

September 2015

Thesis summary

Background: Diagnostic prediction rules are tools to assist clinical decision making and aim to either improve patients' health or provide other benefits without adversely affecting patients. However their uptake in clinical practice has been limited. Possible explanations include uncertainty about their performance compared to and in combination with clinicians' clinical judgment and complexity for use at the bedside.

Objectives: My four aims in this thesis were to:

1. compare the diagnostic performance of diagnostic prediction rules and clinical judgment versus a reference standard;
2. determine the effect of care provided with and without diagnostic prediction rules on patient and process outcomes;
3. derive and validate a prediction rule for the identification of children with serious bacterial infection in primary care, and to determine the accuracy, independent and added value of an inflammatory biomarker, C-reactive protein (CRP). An existing dataset was to be used for the derivation study and study of the added value of CRP;
4. investigate how simplifying a prediction rule affected performance.

Methods: To address aim 1, I completed a systematic review comparing the diagnostic performance of diagnostic prediction rules and clinical judgment against a reference standard. To address aim 2, I completed a systematic review comparing the effect of care provided with and without a diagnostic prediction rule on patient health and process outcomes. To address aim 3, I completed a systematic review of the diagnostic accuracy and independent value of CRP for detecting serious bacterial infection in non-hospitalised children. However, due to the volume of missing data in the dataset sourced to derive the prediction rule, a valid prediction rule could not be derived. To address aim 4, I conducted a study in which an existing prediction rule was simplified using several methods, and the effect on performance assessed.

Results: Existing diagnostic prediction rules were not clearly superior to clinical judgment in terms of diagnostic performance. In some situations prediction rules moved the threshold for diagnosing disease such that fewer patients with disease were missed, but this was at the cost of further investigations in a larger proportion of patients, or vice versa. The findings are limited by the small number and potential biases of the included studies. Diagnostic prediction rules improved symptoms and reduced antibiotic prescribing for sore throat, improved early

discharge and hospitalisations for possible cardiac chest pain and reduced time to therapeutic operations in suspected appendicitis, but did not improve process outcomes in studies of children with fever. Few studies evaluated patient health outcomes and details of study interventions and implementation were infrequently reported. CRP provides moderate diagnostic information for ruling in and ruling out serious bacterial infection in non-hospitalised children and diagnostic information independent of other clinical features. Simplifying a diagnostic prediction rule did not affect overall accuracy, but reduced the proportion of patients classified as low risk and resulted in worse classification.

Summary: In terms of diagnostic performance, existing diagnostic prediction rules do not clearly outperform the judgment of clinicians. However, they may improve patient health and process outcomes in some clinical conditions. C-reactive protein provides useful diagnostic information for children with suspected bacterial infection. Simplification reduced the performance of one diagnostic prediction rule with acceptability of this context dependent.

Declaration and Addendum

This thesis is submitted to Bond University in fulfilment of the requirements of the degree of *Doctor of Philosophy*. This thesis represents my own original work towards this research degree and contains no material which has been previously submitted for a degree or diploma at this University or any other institution, except where due acknowledgement is made.

Sharon Lea Sanders is the sole author of the Introduction and Discussion chapters, Chapter 4 and lead author on all other chapters which are substantially unchanged multi-author papers. The original research work underpinning Chapters 2, 3, 5 and 6 was driven primarily by Sharon Lea Sanders, who managed all aspects of the collaborative research projects and also produced initial, subsequent and final drafts of each manuscript. None of the work submitted in this thesis was carried out before the PhD candidature.

As per university rules, where a substantially unchanged multi-author paper is included in the thesis, a statement appears at the end of the chapter outlining the contributions of all involved, and these statements have been signed by all authors.

Copyright Declaration

I hereby grant to the Bond University or its agents the right to archive and to make available my thesis or dissertation in whole or in part to the University libraries in all forms of media, now and hereafter known.

In regard to the digital version of the thesis I have clearly separated and identified all the parts that have been previously published. Where copyright has been transferred to a publisher I have gained permissions to leave the publications in the digital version when it is placed online in the University's digital repository, e-publications@bond.

I have attained all copyright permissions required for use of third party material and retain all proprietary rights, such as patient rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I have provided a copy of all copyright permission letters with the thesis

Acknowledgements

It has been a privilege to have been able to undertake this interesting and challenging program of research.

I am sincerely grateful to my long suffering Principle and Associate advisors, Professor Jenny Doust and Professor Paul Glasziou. Thank you for your belief in a student endowed with a (probably unhealthy) dose of self-doubt. I deeply relied upon your gentle encouragement, patience and belief in me. Thank you for your interest in my work and for your willingness to share your skill and expertise. I feel tremendously fortunate to have been your student. Thank you also to my Associate advisor Katy Bell, whose enthusiasm and assistance late in the thesis has been invaluable.

I would also like to thank Professor Ian Scott, Associate Professor Adrian Barnett and my colleagues at the Centre for Research in Evidence Based Practice at Bond University who have provided intellectual and other forms of support during this thesis. I would particularly like to thank Professor Chris Del Mar, Associate Professor Elaine Beller, John Rathbone, Ray Moynihan and Rae Thomas for her encouragement and advice during impromptu 'counselling' sessions. My sincere thanks also to Dr Martin Than, Dr Dylan Flaws and Dr Matthew Thompson, for encouraging my work and allowing me to access study data.

I am grateful for the financial support I received during this thesis in the form of an Australian Postgraduate Award, and from funding provided by The Screening and Diagnostic Test Evaluation Program (STEP).

To Ram, Mum, Appa, Amma, Ella , the Triple J's and the rest of my family, I realise this has not been without great sacrifice. Thank you for never, or very rarely, asking why.

Finally, I am very grateful for the opportunity to submit this thesis and sincerely appreciate the time and effort that will go into the examination procedure.

Peer reviewed journal articles arising from this thesis

1. **Sanders S**, Doust J, Glasziou P. A systematic review of studies comparing diagnostic clinical prediction rules with clinical judgment. PLOS One. 2015;10(6):e0128233.
2. **Sanders S**, Barnett A, Correa-Velez I, Coulthard M, Doust J. Systematic review of the diagnostic accuracy of C-reactive protein to detect bacterial infection in nonhospitalized infants and children with fever. The Journal of Pediatrics. 2008 Oct;153(4):570-4.
3. **Sanders S**, Flaws D, Than M, Pickering JW, Doust J, Glasziou P. Simplification of a scoring system maintained overall accuracy but decreased the proportion classified as low risk. Journal of Clinical Epidemiology. 2015 May 14.

Manuscripts submitted and under review

1. **Sanders S**, Rathbone J, Bell K, Glasziou P, Doust J. A systematic review of the effects of diagnostic clinical prediction rules. Submitted to PLOS Medicine 1/9/2015.

Peer reviewed journal articles published that relate to, but did not arise from this thesis

1. Pluddemann A, Wallace E, Bankhead C, Keogh C, Van der Windt D, Lasserson D, Galvin R, Moschetti I, Kearley K, O'Brien K, **Sanders S**, Mallett S, Malanda U, Thompson M, Fahey T, Stevens R. Clinical prediction rules in practice: review of clinical guidelines and survey of GPs. The British journal of general practice: the journal of the Royal College of General Practitioners. 2014 Apr;64(621):e233-42.
2. Than M, Flaws D, **Sanders S**, Doust J, Glasziou P, Kline J, et al. Development and validation of the Emergency Department Assessment of Chest pain Score and 2 hour accelerated diagnostic protocol. Emergency medicine Australasia: EMA. 2014 Feb;26(1):34-44.

3. Mant J, Doust J, Roalfe A, Barton P, Cowie MR, Glasziou P, Mant D, McManus RJ, Holder R, Deeks J, Fletcher K, Qume M, Sohanpal S, **Sanders S**, Hobbs F. Systematic review and individual patient data meta-analysis of diagnosis of heart failure, with modelling of implications of different diagnostic strategies in primary care. *Health Technology Assessment*. 2009 Jul;13(32):1-207.

Table of Contents

Thesis summary.....	ii
Declaration and Addendum	iv
Copyright Declaration	vi
Acknowledgements	viii
Peer reviewed journal articles arising from this thesis	x
Table of Contents	xii
Abbreviations and Acronyms	xvi
List of Tables	xviii
List of Figures	xx
List of boxes	xxii
Chapter 1 Introduction and literature review	1
1.1 Introduction.....	2
1.2 Diagnostic reasoning and error in healthcare	6
1.3 Clinical prediction rules as a strategy for improving diagnostic reasoning and reducing error ..	8
1.4 The derivation, validation and assessment of the impact of clinical prediction rules	15
1.5 Potential harms and unintended consequences of clinical prediction rules.....	23
1.6 Standards of evidence for clinical prediction rules	24
1.7 Methodological standards of published clinical prediction rules	25
1.8 The use of clinical prediction rules	25
1.9 Gaps in the evidence base and research justification	31
1.10 Research aims and thesis overview	33
Chapter 2 The comparative performance of diagnostic prediction rules and clinical judgment.....	37
2.1 Preface to Chapter 2.....	38
2.2 Abstract	39
2.3 Introduction.....	40
2.4 Methods	40
2.5 Results	43
2.6 Discussion	60
Chapter 3 The effects of care provided with and without diagnostic prediction rules on patient and process outcomes	65
3.1 Preface to Chapter 3.....	66
3.2 Abstract	67
3.3 Introduction.....	68

3.4	Methods	69
3.5	Results	73
3.6	Discussion	97
3.7	Conclusion	101
Chapter 4 The identification of serious bacterial infection in children with fever presenting to primary care..... 103		
4.1	Preface to Chapter 4.....	104
4.2	The clinical problem	105
4.3	Strategies to improve the diagnostic management of children with possible serious bacterial infection presenting to primary care	106
4.4	Proposed studies intended to assist clinicians' diagnostic management of children with possible serious bacterial infection presenting to primary care	107
4.5	The research completed, incomplete and research plan modification	109
4.6	The current status of diagnostic prediction rules as a strategy for assisting clinicians' diagnostic management of children with possible serious bacterial infection presenting to primary care	112
Chapter 5 The diagnostic accuracy and independent value of C-reactive protein for detecting bacterial infection in nonhospitalised infants and children with fever..... 117		
5.1	Preface to Chapter 5.....	118
5.2	Abstract	119
5.3	Introduction.....	120
5.4	Methods	120
5.5	Results	122
5.6	Discussion	129
5.7	The current status of C-reactive protein for assisting in the diagnostic management of non-hospitalised children with fever and possible serious infection.....	130
Chapter 6 The effect of simplification of a diagnostic prediction rule presented as a scoring system on performance 133		
6.1	Preface to Chapter 6.....	134
6.2	Abstract	135
6.3	Introduction.....	136
6.4	Methods	137
6.5	Results	140
6.6	Discussion	146
6.7	Conclusion	149
Chapter 7 Discussion and Conclusions 151		
7.1	Preface to Chapter 7.....	152
7.2	Summary of findings.....	153

7.3	Limitations and strengths of the studies and thesis in the context of the wider literature	156
7.4	Implications and recommendations arising from this thesis.....	159
7.5	Conclusion	164
Appendix A Supplementary material from the systematic review of studies comparing diagnostic clinical prediction rules with clinical judgment		165
Appendix B Supplementary material from the systematic review of the effects of diagnostic clinical prediction rules		171
Appendix C Supplementary material from the study of the simplification of a clinical prediction rule.....		185
Reference list		189

Abbreviations and Acronyms

AUROC	Area under the Receiver Operating Characteristic curve
AVDSf	Alveolar dead-space fraction
CI	Confidence interval
CONSORT	Consolidated Standards of Reporting Trials
COPD	Chronic obstructive pulmonary disease
CPR	Clinical prediction rule
CRP	C-reactive protein
CT	Computed tomography
DVT	Deep vein thrombosis
ECG	Electrocardiogram
ED	Emergency department
EDACs	Emergency Department Assessment of Chest Pain Score
ES	Emergency services
FN	False negative
FP	False positive
F/U	Follow-up
HR	Hazard ratio
HTA	Health Technology Assessment program
IP	Inpatient clinic
LR	Likelihood ratio
MD	Mean difference
MACE	Major adverse cardiac event
NMD	Nuclear medicine department
NRI	Net reclassification improvement
OAR	Ottawa Ankle Rules
OPC	Outpatient clinic
OR	Odds ratio
PC	Primary care

PE	Pulmonary embolism
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RADT	Rapid antigen detection test
RCT	Randomised controlled trial
RR	Risk ratio
SBI	Serious bacterial infection
SD	Standard deviation
SDC	Structured data collection
STARD	Standards for the Reporting of Diagnostic accuracy studies
SU	Surgical unit
TP	True positive
US	Ultrasound
UTI	Urinary tract infection
VQ	Ventilation perfusion scan
VL	Vascular laboratory
wNRI	Weighted net reclassification improvement

List of Tables

Table 1.1. Literature informed potential barriers to the adoption of clinical prediction rules in practice	30
Table 2.1. Clinical conditions and study comparisons	46
Table 2.2. Risk of bias and applicability concerns for individual studies included in the review	47
Table 2.3. Characteristics and results of included studies for conditions with >2 studies	53
Table 2.4. Characteristics and results of included studies for conditions with ≤2 studies	56
Table 3.1. Characteristics of the included studies by clinical condition.....	76
Table 3.2. Risk of bias in the included studies	81
Table 3.3. Results of studies of Group A Streptococcus throat infection by outcome	84
Table 3.4. Results of studies of acute appendicitis by outcome	87
Table 3.5. Results of studies of serious bacterial infection in children with fever	89
Table 3.6. Results of studies of ankle or mid-foot fracture by outcome	90
Table 3.7. Results of studies of acute coronary syndrome by outcome	91
Table 3.8. Results of single studies of different clinical conditions	93
Table 3.9. Minimum required elements for reporting of diagnostic strategies and implementation methods	95
Table 4.1. Clinical prediction rules comprised of predictors from history or physical examination for serious bacterial infection overall or specific serious bacterial infections in children (published to end 2007).....	108
Table 4.2 Clinical prediction rules comprised of predictors from the history or physical examination (with or without laboratory predictors) for serious bacterial infection overall or specific serious bacterial infections in children derived and or evaluated since 2007	113
Table 5.1. Details of studies investigating the accuracy of C-reactive protein (CRP) to differentiate serious bacterial infection from benign bacterial/nonbacterial infection and bacterial from nonbacterial/viral infection	123
Table 5.2. Quality of the included studies with useable 2 x 2 data according to QUADAS criteria	125
Table 5.3. Performance of C-reactive protein (CRP) for the detection of serious bacterial infection and bacterial infection.....	128
Table 6.1. Original EDACS and simplified scores.....	139
Table 6.2. Characteristics of the derivation and validation cohorts	141

Table 6.3. Area under the receiver operator characteristic curve (AUROC) for the original and simplified versions of the score and sensitivities and specificities when the score is used in conjunction with ECG and c-troponin tests	143
Table 6.4. Proportion of patients assigned to low and high risk categories by the scores when used in combination with ECG and c-troponin tests by event in the validation dataset (n=909).....	143
Table 6.5. Changes in classification with simplified scores when used in conjunction with ECG and c-troponin tests.....	144

List of Figures

Figure 1.1 Studies of clinical prediction research (including predictor finding studies and clinical prediction rule studies) in PubMed published between 1994 and 2014 as a fraction of the total number of studies in PubMed	10
Figure 1.2. Simplified test-treatment pathway showing the components of patient management that can influence patient health	11
Figure 1.3 The threshold approach to diagnosis	12
Figure 1.4. Stages in the development of a clinical prediction rule	15
Figure 2.1. PRISMA flow diagram of the article selection process	45
Figure 2.2. Summary QUADAS-2 risk of bias and applicability judgments	47
Figure 2.3. Results of the included studies for conditions with >2 studies.....	58
Figure 2.4. Results of the included studies for conditions with ≤ 2 studies.....	59
Figure 3.1. Study flow diagram	74
Figure 3.2. Meta-analysis of Group A Streptococcus throat infection studies for the outcome antibiotic prescriptions	84
Figure 3.3. Meta-analysis of acute appendicitis studies for the outcome unnecessary appendectomies	88
Figure 5.1. Results of studies estimating the sensitivity and specificity of C-reactive protein for the detection of serious bacterial infection and bacterial infection.	127
Figure 6.1. Comparison of original and simplified scores used in combination with ECG and c-troponin tests in validation dataset (n=909)	145

List of boxes

Box 1-1. A clinical prediction rule for chronic obstructive pulmonary disease (COPD)	4
Box 1-2. A clinical prediction rule (Ottawa Ankle Rules) providing a testing recommendation...	5
Box 1-3. A clinical prediction rule (The Modified Centor Score) providing a testing and treatment recommendation.....	5
Box 1-4. Multivariable logistic regression model for the diagnosis of pulmonary embolism showing the estimated regression coefficients which reflect the relative 'weight' of each predictor.....	13
Box 2-1 Outcomes of the review.....	43

Chapter 1 Introduction and literature review

1.1 Introduction

A clinicians' ability to diagnose accurately is central in determining an individuals' prognosis and providing timely and appropriate treatment. However, making a diagnosis is often challenging and uncertainty an error a common feature of the diagnostic process. Clinical prediction rules are tools developed to assist clinicians' clinical reasoning and decision making. (1, 2) Clinical prediction rules designed for use in the diagnostic setting provide an estimate of the probability of an outcome being present and/or suggest a course of clinical action for an individual based on the underlying probability estimate using data on multiple individual level characteristics. There is a belief or expectation among proponents of these tools, that these tools will enhance diagnostic reasoning and decision making, and ultimately improve patient health or provide other benefits without causing harm.

1.1.1 The terminology of clinical prediction rules

The term 'clinical prediction rule' is generally used to refer to variably presented and labelled prediction tools that convert a combination of predictor values from an individual (such as age or gender, symptoms or findings from examination or laboratory or imaging investigations) to an estimate of the probability that a certain condition or disease is present (diagnosis) or will occur in the future (prognosis). When the clinical prediction rule provides an estimate of the probability that a certain condition or disease is present at the moment of prediction, and is developed for used in individuals suspected of having that condition, it is a diagnostic clinical prediction rule. When the clinical prediction rule estimates the probability of a particular outcome or event occurring in a certain time period in the future, and is intended for use in individuals at risk of developing that outcome, it is a prognostic clinical prediction rule. In earlier eras, clinical prediction rules were predominantly developed based on expert opinion, but the majority of contemporary clinical prediction rules are created by applying multivariable statistical techniques to data from patients in whom the outcome of interest is known.

Clinical prediction rules are known by an array of synonymous terms including prediction tools or guides, decision rules, algorithms or risk scores. They may also be termed 'decision aids' or 'decision support systems'. However, while decision aids and decision support systems may incorporate a clinical prediction rule, they are complex interventions that are not synonymous with clinical prediction rules (3, 4) Though there is no agreement among developers and researchers in the area about the most appropriate term used to label clinical prediction tools, there is a desire to differentiate between and differentially label prediction tools that provide guidance to clinicians in the form of a probability estimate that reflects the continuum between absolute certainty ($P_i=1$) and certified impossibility ($P_i=0$) as prediction 'models' (see

Box 1-1), and those that, through the application of a decision threshold to the probability estimate, classify an individual into a risk group (i.e. high or low risk or positive or negative for disease) as prediction 'rules'. (5, 6) There is also a desire to distinguish between prediction tools that provide a the user with a probability estimate only, and those that explicitly recommend clinical actions. The former, preferentially labelled 'prediction' rules, intend to assist clinicians without telling them what to do and assume that accurate predictions will improve clinical decisions. The later, known as 'decision' rules aim to effect clinical decisions directly. (5) The recommendations provided by diagnostic 'decision' rules may relate to further diagnostic testing or treatment or both. For instance, the Ottawa Ankle Rules, (7) when used in the assessment of acute ankle injury, identify a group of patients in whom fracture is very unlikely and x-ray unnecessary (Box 1-2). The Modified Centor Score, (8) when used in patients with a sore throat, provides three possible recommendations on further diagnostic testing and treatment (Box 1-3).

1.1.2 The focus of and terminology used in this thesis

Prediction tools developed for use in the diagnostic setting and derived from multivariable statistical analyses are the focus of this thesis. Such tools provide an estimate of the probability of the presence of an outcome (and/or recommend a clinical course of action) for an individual *at the moment of prediction* and are developed for use in individuals suspected of having that condition. (1) The terminology to be used in this thesis is outlined below. It is provided as a guide to facilitate reading of the thesis rather than as a definitive lexicon.

Diagnostic prediction rule: This term is used throughout the introduction, Chapter 4 and discussion section of the thesis to refer to prediction tools developed for use in the diagnostic setting that provide *either* an estimate of the probability of the presence of the disease of interest (so called 'prediction' rules) and/or recommend a course of clinical action (so called 'decision' rules). The information presented in the introduction and discussion of the thesis relates, as far as possible, to diagnostic prediction rules specifically. However, where the literature presented does not differentiate between diagnostic and prognostic prediction rules (as in the section on barriers to use of clinical prediction rules); the term 'clinical prediction rule' is used.

Clinical prediction rule: In the introduction and discussion sections of this thesis, the term clinical prediction rule refers to prediction tools that may be used in the diagnostic or prognostic setting *and* that *either* provide a probability estimate and/or recommend a course of clinical action. In Chapters 2, 3, and 6, which present published manuscripts, the term

clinical prediction rule or CPR is used for brevity as preferred by journal editors, though when used, the terms refer specifically to diagnostic clinical prediction rules.

1.1.3 Applications of diagnostic prediction rules

In the diagnostic setting, the probability estimate or risk classification provided by a diagnostic prediction rule may inform clinicians' decisions as to whether a particular condition can be safely ruled out or ruled in, or be used for shared decision making with patients. When the diagnostic prediction rule provides a recommended course of action, this may be used to assist clinicians' decisions on whether or not to order more invasive or costly diagnostic tests, or to identify patients who may benefit from referral. (1, 2, 9-11)

Box 1-1. A clinical prediction rule for chronic obstructive pulmonary disease (COPD)

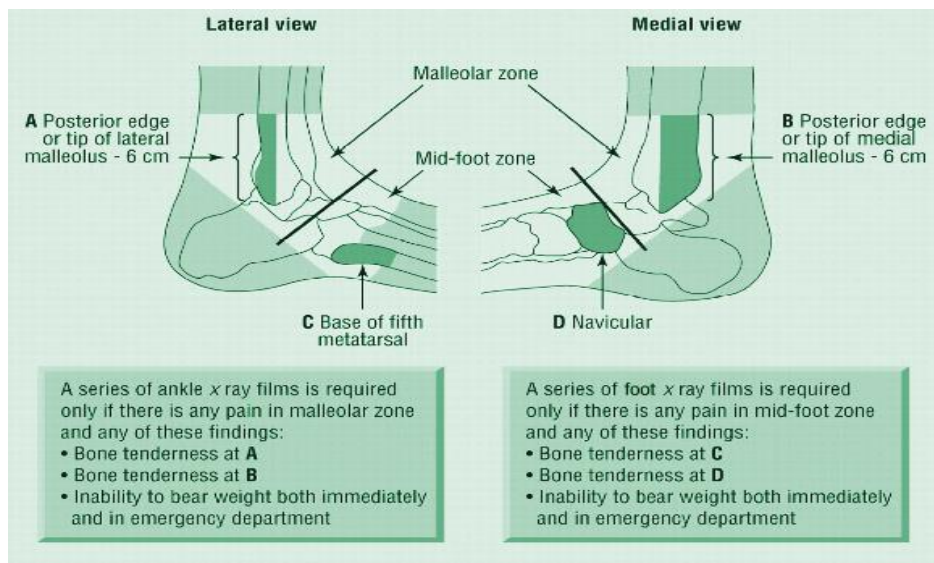
Probability of COPD presence = $1/(1 + \exp - [-5.6 + 0.2 \times \text{every 5 years above age of 50} + 0.6 \times \text{male sex} + 1.0 \times \text{current smoking} + 0.7 \times >20 \text{ pack years} + 0.6 \times \text{cardiovascular disease} + 0.6 \times \text{complaints of wheezing} + 0.7 \times \text{diminished breath sounds}])$

For a 62 year old male who complains of wheezing the predicted probability (pCOPD) is calculated as follows;

$$1/(1 + \exp - [-5.6 + 0.2 \times 2 + 0.6 \times 1 + 1.0 \times 0 + 0.7 \times 0 + 0.6 \times 0 + 0.6 \times 1 + 0.7 \times 0])$$

pCOPD = 2 %

Republished with permission of the Royal College of General Practitioners, from Does a decision aid help physicians to detect chronic obstructive pulmonary disease?, Broekhuizen et al., 61(591), 2011. (12); permission conveyed through Copyright Clearance Centre, Inc.

Box 1-2. A clinical prediction rule (Ottawa Ankle Rules) providing a testing recommendation

Reprinted from the Annals of Emergency Medicine, 21(4) Stiell et al., A study to develop clinical decision rules for the use of radiography in acute ankle injuries, 384-90, 1992 (7), with permission from Elsevier.

Box 1-3. A clinical prediction rule (The Modified Centor Score) providing a testing and treatment recommendation

Criteria	Points
Temperature $>38^{\circ}\text{C}$	1
Absence of Cough	1
Swollen, Tender Anterior Cervical Nodes	1
Tonsillar Swelling or Exudate	1
Age	
3-14 Years	1
15-44 Years	0
45 Years or Older	-1

Score	Risk of Streptococcal Infection ^{8,9}	Suggested Management
≤ 0	1%-2.5%	No Further Testing or Antibiotic
1	5%-10%	
2	11%-17%	Culture All;
3	28%-35%	Antibiotics Only for Positive Culture Results
≥ 4	51%-53%	Treat Empirically With Antibiotics and/or Culture

Reprinted from McIsaac et al., A clinical score to reduce unnecessary antibiotic use in patients with sore throat, Figure 2, Canadian Medical Association Journal (1985, 158(1), 75-83. (8) © Canadian Medical Association (1985). This work is protected by copyright and the making of this copy was with the permission of the Canadian Medical Association Journal (www.cmaj.ca) and Access Copyright. Any alteration of its content or further copying in any form whatsoever is strictly prohibited unless otherwise permitted by law.

1.2 Diagnostic reasoning and error in healthcare

The ability to transform medical data into an actionable diagnosis is the most critical of clinicians' skills (13) and is central to providing timely and appropriate treatment. It is of great importance to clinicians and patients alike. If a correct diagnosis is not made or is delayed, effective treatment may be withheld or patients may undergo inappropriate medical treatment, or they may receive inaccurate information about their prognosis.

1.2.1 Clinical reasoning and diagnosis

According to the prevailing model of clinical reasoning, clinicians' decisions about diagnoses are characterised by both fast, unconscious (System 1), and slow, analytical conscious (System 2) processes. (13) System 1 processing, also known as non-analytical processing, is characterised as an automatic and intuitive, largely heuristic process that allows clinicians to formulate diagnostic hypotheses efficiently and rapidly. System 2 processes are recognised as analytical, logical, deliberate processes of carefully and systematically gathering and weighting information. The systems are consistent with the recognised reasoning strategies of pattern recognition and hypothetic-deductive reasoning in their *modus operandi* and have been shown to be congruent with the reasoning processes of clinicians from several specialities. (14-16) It is believed both systems are jointly involved in reasoning, with valance towards one system or another depending upon time constraints, the complexity of the situation, the nature of the problem (ambiguous, non-routine or ill-defined problems) and the context of uncertainty. (17)

1.2.2 Diagnostic error

Despite the importance of accurate diagnosis, diagnostic errors occur at an appreciable rate. (18) In a recent review of studies, diagnostic errors (defined as unintentionally delayed, wrong or missed diagnoses as judged from the eventual appreciation of more definitive information (19)) were found to occur at a rate of approximately 5% in perceptual specialties (for example radiology where diagnosis is made based on the perception of an image) and up to 15% in other settings requiring more data gathering and synthesis (for example primary care and emergency departments). (20) Although not all diagnostic errors translate into harm, a substantial number are associated with preventable morbidity and mortality (21) with diagnostic errors reported to result in death or disability twice as often as other types of medical error. (22)

Causes of diagnostic error have been found to include one or more of the following: clinician cognitive factors (clinicians perceptual and thought processes) or faulty knowledge or skills; systems factors (organisational, technical and equipment problems); and patient factors

(communication practices and variability in clinical presentation). (19, 23) In a study to determine the contribution of system related and cognitive aspects to diagnostic error, diagnostic error was commonly multifactorial in origin with both system and cognitive factors the cause of error in 46% of cases. Cognitive factors were found to contribute to diagnostic error in 74% of cases (in 28% of cases it was the sole cause of error) and system related factors in 65%, while errors arising from faulty or inadequate knowledge were uncommon. (19)

The cognitive failings that lead to diagnostic error have typically been attributed to cognitive biases associated with the non-analytical, intuitive (system 1) thinking, though they have also been found to be associated with analytical (system 2) reasoning. (24-26) A substantial number of cognitive biases that may affect decision making in the non-medical world have been identified and described and applied to the domain of medical practice. (27, 28) In one of the most extensive studies of the mechanisms of diagnostic errors, errors arising from flawed processing of the information gathered were more common than errors caused by knowledge gaps. (19) In this study, the most common information synthesis error was premature closure; the tendency to stop considering other possibilities after reaching a diagnosis. Other common causes of error were faulty context generation, faulty perception (for example incorrect reading of x-rays) and failed use of heuristics. (19) It is noted however, that the currently used research methods for examining the causes of diagnostic error are limited for examining the causal relationship between diagnostic reasoning and error. (29) As such, naturalistic and experimental research is being undertaken to determine the connection between diagnostic reasoning processes and diagnostic errors. (30, 31)

Other inherent aspects of the diagnostic task, primarily the amount and mutual dependence of information collected during the diagnostic workup, also adds significant challenges to achieving accurate diagnosis and optimal decisions. Diagnostic reasoning often involves the collection of a substantial amount of information obtained from the patient history and physical examination and application of laboratory and imaging tests. In principle, the clinician implicitly integrates the information received into a judgment regarding the probability of the suspected or differential diagnosis. (32) However, this information is often received in an idiosyncratic manner and may simply be too much for clinicians to process (so-called bounded rationality). Further, the information obtained is to varying degrees overlapping or mutually dependent. Simply because diagnostic tests contribute either directly or indirectly to the cause or the result of the same underlying disease, they are likely to be correlated and provide to some extent the same information. For example, creatine kinase and troponin are blood tests that may be used in the assessment of chest pain. Both tests measure enzymes found in the

blood that are released when there is damage to the heart muscle cells. Because the enzymes measured by these tests occur through a related pathological mechanism there is a degree of dependency between the tests. Mutual dependency between test results means that the diagnostic potential of a test is conditional upon the information obtained from previous tests, (32) that is, the diagnostic value of troponin changes when creatine kinase is also used.

Accounting for the mutual dependency between test results and reliably judging the relative contribution or true diagnostic value of the multiple symptoms, signs and test results for a diagnosis, presents a formidable obstacle for clinicians. Studies of novice and experienced clinicians have shown that both have difficulties in recognising the diagnostic value of information, leading them to use non-discriminatory findings - signs or symptoms that are not able to accurately validate a diagnosis, to support a diagnosis. (33, 34) A clinicians' understanding of the diagnosticity of clinical information arises primarily from clinical experience. However, clinical experience is fallible:

If our case numbers were truly vast (hundreds, if not thousands), if the spectrum of disease we had seen was sufficiently wide and representative, if we had used diagnostic criteria consistently over time, if we had searched for each clinical finding equally diligently in every patient, and if our memories were perfect, then perhaps our library of remembered cases would allow us to accurately estimate the frequencies of clinical manifestations of that disease and interpret them properly (35 p.A12)

Consequently;

...Short of exposure to truly representative samples, it may be difficult, if not impossible to determine relations among multiple pieces of diagnostic information and to distinguish what is useful from what is useless (36 p.1671)

1.3 Clinical prediction rules as a strategy for improving diagnostic reasoning and reducing error

As part of a greater patient safety focus in healthcare over the last decade, there has been heightened awareness of the extent and implication of diagnostic errors and increasing study aimed at determining the incidence and etiology of these errors. (37, 38) Though recent research in this area suggests that diagnostic errors have both cognitive and systems origins, (19) the prevailing belief has been that diagnostic errors are principally the result of flaws in

the way clinicians think; that clinical judgment is suboptimal and flawed for making diagnosis and good clinical decisions.

As such, strategies for minimising the cognitive shortcomings of clinicians have been pursued.

(39) Diagnostic clinical prediction rules are one such strategy. Based on decades of research showing the superiority of statistical models over clinical judgment as alternate methods for integrating clinical data (36), there is a belief or expectation that prediction rules, through the provision of ‘accurate and objective’ estimates of probability, will lead to improved accuracy of diagnosis and subsequent clinical decisions. These sentiments are reflected in the following statements found in key multivariable prediction rule methodological and position papers:

“Clinical prediction rules ... are formulated to improve the efficiency and accuracy of physicians judgments” (40 p.797)

“Clinical decision rules attempt to formally test, simplify and increase the accuracy of clinicians diagnostic and prognostic assessments” (41 p.79)

“Clinical prediction rules provide powerful tools to improve clinical decision making...” (5 p.207)

“Prediction models are increasingly used to complement clinical reasoning and decision making in modern medicine” (42 p.683)

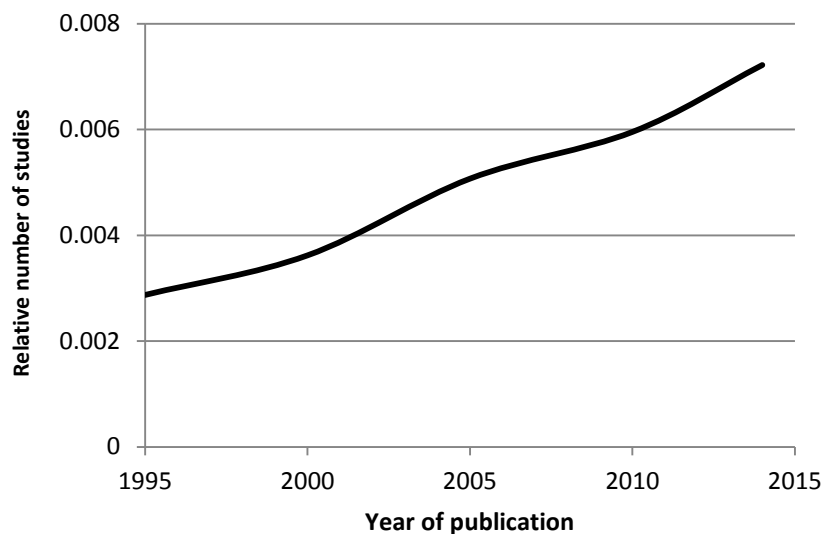
“Prediction models are being developed... with the aim to assist doctors and individuals in estimating probabilities and potentially influence their decision making” (1 p.w1)

Though rarely articulated in these documents, the underlying rationale for the development of diagnostic prediction rules is the expectation and unproven assumption that altered clinical decisions arising from their use will ultimately lead to improved patient outcomes or will confer other benefits (such as reduced resource use) without adversely affecting patients. Also implicit in these statements is the supposition that the two methods of judgment, clinical (clinical judgment) and statistical (clinical prediction rules) are used in combination, and that this approach is superior to either the statistical model (prediction rule) or clinical judgment alone.

These expectations have led to a tremendous increase in the number of published articles relating to clinical prediction research over the last decade. ‘Prediction research’ comprises predictor finding studies (studies which aim to discover which predictors out of a number of candidate predictors independently contribute to the prediction of an outcome), clinical prediction rule studies (studies deriving or validating a multivariable clinical prediction rule) and impact studies (studies which quantify the effect of a clinical prediction rule). This increase is evident in Figure 1.1 which shows the number of prediction research studies (a total of

155,783 in 2014) published in PubMed (identified using a suggested search strategy for retrieving prediction research (43)) as a fraction of the total number of studies indexed in PubMed over the last decade. In relation to clinical prediction rules specifically, reviews have shown the increasing number of publications reporting their development over time (44, 45). In addition, clinical prediction rules are increasingly included in clinical practice guidelines and recommended for use by national bodies (46-48). For every article describing the development of a prediction model that is published, there may be half as many again, that remain unpublished (49).

Figure 1.1 Studies of clinical prediction research (including predictor finding studies and clinical prediction rule studies) in PubMed published between 1994 and 2014 as a fraction of the total number of studies in PubMed

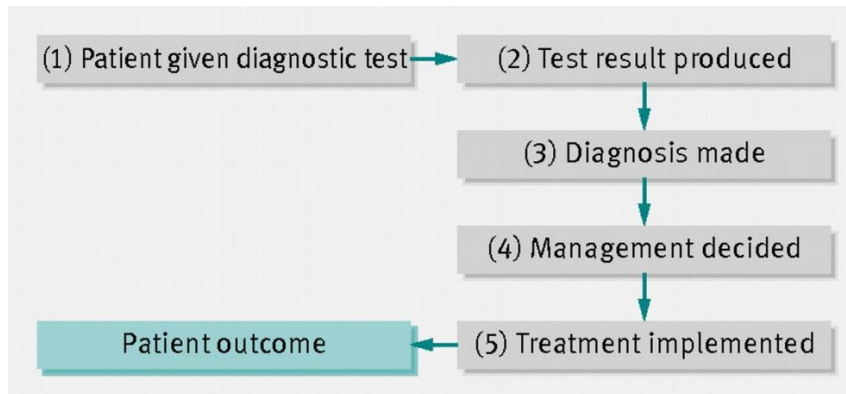


1.3.1 How clinical prediction rules may alter clinical decision making and influence patient outcomes.

The mechanisms by which clinical prediction rules may alter decisions and health outcomes are rarely articulated in studies describing their development. However, based on a framework of mechanisms through which a diagnostic test can cause changes to patient health, (50) the means by which a diagnostic prediction rule may conceivably influence clinical decisions and patient outcomes can be explored.

The basic relationship between diagnostic tests such as diagnostic prediction rules and patient outcomes may be understood as a series of intermediate steps occurring between the two. This relationship is depicted in a simplified test-treat pathway (Figure 1.2). (50) Changes to any aspect of the pathway caused by the introduction of a diagnostic prediction rule could induce changes in patient outcomes.

Figure 1.2. Simplified test-treatment pathway showing the components of patient management that can influence patient health



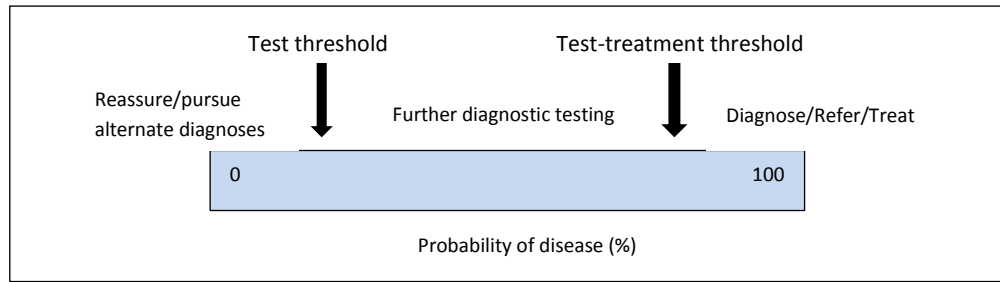
Reproduced from The BMJ, Ferrente di Ruffano et al., 344, e686, 2012 (50), with permission from BMJ Publishing Group Ltd.

1.3.1.1 Altering clinical decisions and actions

The most widely recognised means for a diagnostic test such as a clinical prediction rule to affect downstream patient health is through its capacity to change decision making and management guided by the test result. (51) Depending upon the information provided by a diagnostic prediction rule (a probability estimate or a management recommendation), it may conceivably act to alter testing or treatment decisions directly or via a pathway of improved diagnostic accuracy.

1.3.1.1.1 Altering clinical decisions and actions via changes to the diagnostic accuracy of clinical judgment

Both the outcome of a prediction model or rule (i.e. the probability estimate or risk classification) and the actual process of using the rule may alter the clinicians' ability to produce classifying information more accurately. Within the framework of probabilistic reasoning, in which a clinician continually revises their estimate of the probability of the disorder as clinical information is obtained, the estimated probability provided by a diagnostic prediction rule may either move a clinicians probability estimate above the test- treatment threshold in which case a diagnosis is made and treatment can be started, or below the test threshold where the probability of the disease being present is so low that either the individual can be reassured the disease is not present, an alternative diagnoses be pursued, or a watchful waiting practice adopted (Figure 1.3). (52, 53)

Figure 1.3 The threshold approach to diagnosis

Alternatively, or in addition, by specifying the data required to derive output from the model, a clinician's attention is focused on factors from the history or examination that are known to be predictive of the disease of interest and possibly to discounting clinical variables that are less predictive. This 'honing of their gestalt', in itself, may have an effect on accuracy. (54)

For a multivariable diagnostic prediction rule to enhance the diagnostic accuracy of a clinician, the prediction model or rule itself must provide more accurate probability estimates or have greater ability to distinguish between individuals with and without disease than clinical judgment alone. There are several reasons to believe that the outputs of prediction rules would be more accurate than clinical judgment. Firstly, when derived using multivariable regression modelling techniques applied to data from patients in whom the outcome is known, diagnostic prediction rules specify the combinations of predictors (information from the history, examination or test results) that, in view of other test results, have true predictive value for the outcome of interest. That is, they account for the mutual dependencies (correlation) between the different test results and ensure only those pieces of diagnostic information from the history, examination or test results that independently contribute to the estimation of the probability of disease presence are included and information that has no predictive power is ignored. Secondly, when multivariable logistic regression is used to derive a prediction model, the predictors are 'weighted' in accordance with their contribution to the estimation of the probability of disease presence. The estimated regression coefficients reflect the relative 'weight' of each predictor when mutually adjusted for the other predictors in the model, thus quantifying the independent contribution of each predictor to the outcome probability or risk estimation. (42) This is illustrated in Box 1-4 below.

Box 1-4. Multivariable logistic regression model for the diagnosis of pulmonary embolism showing the estimated regression coefficients which reflect the relative 'weight' of each predictor.

Predictor	Regression coefficient (SE)	OR (95% CI)	P-value
Intercept	-3.75 (0.34)	-	-
Clinical signs and symptoms DVT	1.93 (0.33)	6.9 (3.6-13.2)	<0.01
PE most likely diagnosis	1.32 (0.34)	3.8 (1.9-7.3)	<0.01
Heart rate > 100 beats min ⁻¹	0.90 (0.31)	2.4 (1.3-4.5)	<0.01
Recent immobilization or surgery	0.71 (0.32)	2.0 (1.1-3.8)	0.03
Previous DVT or PE	0.91 (0.34)	2.5 (1.3-4.8)	<0.01

DVT, deep venous thrombosis; PE, pulmonary embolism; OR, odds ratio; CI, confidence interval; SE standard error. The intercept reflects the baseline risk. The regression coefficient reflects the relative weight per predictor. The exponent of a regression coefficient yields the odds ratio (OR) of the predictor. An OR of 2.0 for the predictor 'recent immobilization or surgery' indicates that the odds of having PE in a patient suspected of PE is twice as high if the predictor is present, compared with a situation in which the predictor is absent, all other predictors kept constant.

Reproduced from Diagnostic and prognostic prediction models, Hendriksen et al., *Journal of Thrombosis and Haemostasis*, 11(Suppl 1). (10) Copyright © 2013 John Wiley and Sons.

The comparative accuracy of clinical judgment and so called, 'statistical methods' of prediction, in which predictions are based on empirically established relations between patient data and the condition to be predicted, has been extensively studied and debated. (36, 55-59) With few exceptions, studies and reviews published from the 1950s comparing the two methods of prediction from across a wide range of domains including finance, education, psychology and medicine, have concluded that statistical prediction methods are more accurate than clinical procedures. (36, 55, 58) This has been the case for comparisons where judgments are based on the same data (and are thus comparing the interpretive or data combination ability of the methods), in comparisons where judges have access to preferred sources of information (comparing the clinical judges collection and interpretation of information with the statistical method), when the statistical models are simple linear classification rules and when clinical judges have access to the statistical model. (36) However, this conclusion has been challenged by a more recent body of experimental work on heuristics, proposed as descriptive models of human judgment by 'real minds under the constraints of limited knowledge and time' which has found heuristics to be as accurate or, on occasions, more accurate than statistical models. (56)

1.3.1.1.2 Altering clinicians decisions and actions directly

Aside from altering clinical decision via a process of improved diagnostic accuracy, diagnostic prediction rules which provide specific testing and treatment recommendations may act directly to modify clinicians' decisions. Conceivably, the recommendations of a prediction rule may alter a decision that is already made by the clinician, or guide the clinician to a decision different to one they would have arrived at without the input of the prediction rule. Whether

either of these could occur would likely depend on the clinician's confidence in the tool providing the recommendation.

1.3.1.2 *Altering the timing of diagnosis or clinical decisions*

Patient outcomes may improve if tests are undertaken earlier or produce results more quickly, triggering an earlier diagnosis or treatment. (50) Prediction rules comprised of few easily obtainable highly diagnostic indicators from the patients history or physical examination may provide results earlier than the alternative test or testing strategy. For example, a diagnostic strategy for the evaluation of chest pain patients comprised of a prediction model based on features from the history and physical examination and NT-proBNP test, may allow important information to be available more quickly than a departmental strategy based on exercise testing that is often not readily available and for which the patient may have to wait. (60)

1.3.1.3 *Altering the impact of the test process*

Diagnostic prediction models and rules may also influence health outcomes independently of subsequent diagnostic or treatment decisions by reducing harms arising from the testing procedure itself. For instance, among children with suspected serious infection, a diagnostic strategy comprised of a prediction rule (composed of elements from history and examination) and simple point of care test may confer direct physical benefits due to the avoidance of more invasive testing that may occur with the alternative method of investigation.

1.3.1.4 *Other mechanisms for altering clinical decisions or outcomes*

Diagnostic tests may also alter patient outcomes through their influence on the patients' perception and experience of the testing process and response to their test result. (50, 61) For instance, patients may feel less confident in the skill of a clinician who utilises a clinical prediction rule or may feel that the investigation has been less thorough. Impressions of thoroughness may encourage improvements in perceptions of health status and lead to changes in patients' behaviours including adherence to medical advice which will affect health outcomes. However, patients' reactions to test results and diagnosis can be unexpected and difficult to predict and can have both beneficial and negative effects. For example, use of a prediction rule may influence the rate of referral or diagnosis. If serious disease is ruled out more promptly psychological benefits may ensue. On the other hand, confirming the presence of disease may increase anxiety and stress which affects aspects of mental health and social and physical wellbeing. (61)

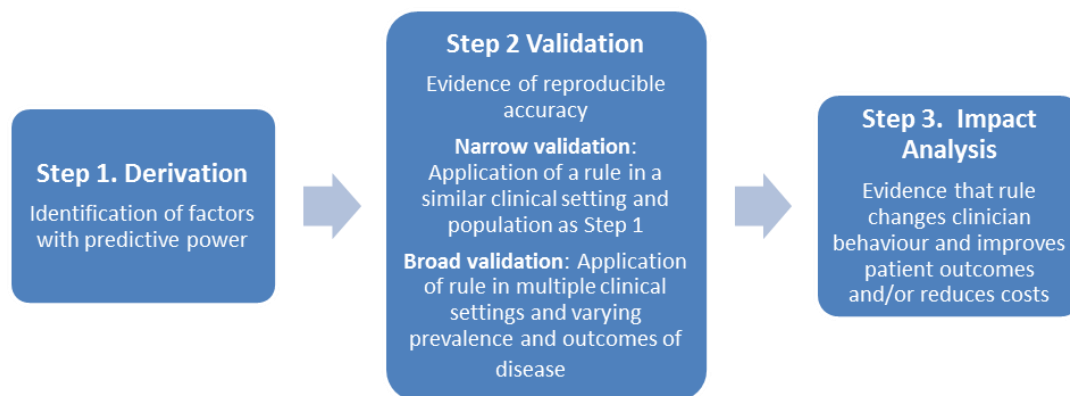
To summarise, diagnostic prediction rules may theoretically affect patient health outcomes in numerous complex ways. In addition to direct effects arising from the attributes of the prediction rule itself, diagnostic prediction rules may alter clinical decisions directly or through

a process of improved accuracy. In another pathway, prediction rules may affect the timeframes of diagnoses and or decisions and actions leading to changes in patient health.

1.4 The derivation, validation and assessment of the impact of clinical prediction rules

Conventionally, multivariable clinical prediction tools are developed in a multistep process involving 3 phases: derivation; validation; and analysis of impact (Figure 1.4). Each phase has a defined purpose and detailed methodological criteria that can be applied *mutatis mutandis* to both prognostic and diagnostic prediction rules have been published. (1, 2, 42, 62-64) The following description of each phase is deliberately brief, with detail provided only on methodological aspects relevant to the remainder of the thesis.

Figure 1.4. Stages in the development of a clinical prediction rule



1.4.1 Stage 1 –Derivation and internal validation of a clinical prediction rule

Derivation of a multivariable diagnostic prediction model involves establishing the cross-sectional relationship between predictor variables (symptoms, signs or diagnostic tests) and the presence or absence of the target condition of interest. This is typically achieved through the application of statistical techniques to data obtained from a study of individuals suspected of having the disease of interest in which the outcome of interest is reported. (65) The resulting model is a mathematical function which relates the presence or absence of the outcome of interest to a set of predictors. To avoid the disease status of some individuals changing, and to maintain the cross sectional relationship of interest, there should be minimal delay between measurement of the predictors and the outcome, and no treatments should be started within this period. In situations where the prevalence of the outcome of interest is low, or the cost of measuring predictors is high, an alternative design may be used where the reference standard is performed first, with all the individuals with the target condition, and a random sample of individuals without the target condition included. (1) It has been observed

that many studies reporting the derivation of a CPR use data from studies originally designed and conducted for a purpose other than the derivation of a CPR. (1) However, the predictive accuracy and ability of such models may be affected due to poorer data quality and unmeasured predictors. (66, 67)

The most common statistical techniques used for diagnostic outcomes (which are usually binary), are multivariable logistic regression and binary recursive partitioning. Classification and regression trees (CART) are a form of binary recursive partitioning in which the study population is progressively divided into subpopulations including only patients with that particular outcome. Both techniques have strengths and weaknesses and one or the other may be preferred in certain situations. (68, 69) In terms of comparative predictive ability, the results of research directly comparing the methods are conflicting. In some studies the predictive ability (as measured by the area under the ROC curve) of logistic regression is comparable to CART (70, 71) and in other comparisons it is superior. (68, 72) Recently, it has been suggested that both approaches can be used together, contributing different advantages. (69) Other, less commonly used statistical methods for deriving diagnostic prediction rules include discriminant analysis, genetic programming and neural networks. Regardless of the statistical technique used to derive the model, internal validation is advocated to estimate the potential for overfitting and optimism in model performance. (66) The preferred statistical method for internal validation is bootstrapping, which aims to mimic random sampling from the source population and yields an average estimate of the amount of overfitting or optimism and adjusts the model accordingly. (1)

During clinical prediction rule derivation, the selection of predictor variables (variables that are chosen to be studied for their predictive performance) is split into two components; 1) the selection of predictors for inclusion in the multivariable analysis (predictor pre-selection), and 2) selection during the multivariable modelling. The use of different predictor selection methods can yield different models and there is no consensus about the best approach. (42) Some form of predictor pre-selection is commonly required because there are usually more predictors available than the investigator can analyse or include in the final prediction model. A common approach for selecting variables for inclusion in the multivariable modelling is based on the association of the candidate predictors with the outcome. However, pre-selection based on univariable significance testing carries a great risk of predictor selection bias, and this method is not recommended. (67) Alternatively, preselection may be achieved a priori by critical consideration of relevant literature or prior knowledge, by combining similar predictors into a single one, or by excluding predictors that are highly correlated with others.

(10) Excluding predictors that are difficult to measure or have high intraobserver variability because this will influence the predictive ability of the model when applied in other individuals is another option. Another consideration for predictor selection before statistical modelling is the context in which the prediction rule will be used. In some settings certain tests are not practical (e.g. high burden or cost) or are unavailable, and these candidate predictors can be excluded from further consideration. In primary care practice, for example, predictor selection may be limited to only predictors that can be practically obtained in that setting such as items from the history or physical examination. In this case, a series of more complex models can sometimes be developed with and without predictors that may or may not be available in that setting (e.g. point of care C-reactive protein).

Just as the method of selecting predictors for inclusion in the multivariable modelling can contribute to optimistic and biased models, so can the method used to select predictors during multivariable modelling. Two of the most commonly used methods are the 'full model approach' and the 'predictor selection approach'. In the full model approach, all a priori selected predictors are included in the multivariable analyses and no further predictor selection is used. This avoids predictor selection bias but this approach requires substantive prior knowledge about the most promising candidate predictors, which is not always straightforward. (2) In the predictor selection approach, candidate predictors that do not contribute usefully in the multivariable model are removed. Backward elimination starts with all candidate predictors in the multivariable model and runs a sequence of tests to remove or keep variables in the model based on a predefined nominal significance level for variable exclusion. The forward selection approach starts with an empty model, and predictors are sequentially added until a pre-specified stopping rule is satisfied. Backward elimination is generally preferred because all correlations between predictors are considered in the modelling procedure (1), however with either method, the criterion for predictors to be selected for inclusion in the model is critical. Possible criteria for predictor inclusion include the use of a nominal p value, the Akaike or Bayesian Information Criterion, or using a change in the models c-index (see the following section). The choice of a relatively small significance level for predictor selection (e.g. $p < 0.05$) generates models with fewer predictors, but increases the chance of missing potentially important predictors, while larger levels (e.g. $p < 0.20$) increase the risk of selecting less important predictors. In both cases, overfitting may arise, particularly in small datasets.(2, 42) The use of the Akaike Information Criterion is the preferred alternative because it accounts for model fit while penalising for the number of parameters being estimated. (2)

The final step in model derivation is consideration of model presentation, and various graphical and simplified formats may be used to present different or the same prediction models and rules. The format selected by developers and implementers to present prediction models and rules may vary according to the desired user friendliness or simplicity of the tool, the information that is to be provided (a probability estimate or a risk classification), the intended audience (with some clinical areas routinely using and preferring certain graphical representations) and whether the model or rule is to be used as a paper based or computerised tool. When used as a computerised tool, the prediction model or rule may be a simple web or non-web based calculator, or may be part of a computerised decision support system integrated with the electronic patient record. Scoring systems are a common simplified presentation format of an underlying regression model aimed to facilitate use of the model in practice. When presented as scoring systems, predictors included in the model are assigned a point value with the points summed to give a total score that is related to absolute outcome probabilities. Assignment of an integer point value to each predictor may be achieved by rounding the regression coefficients, such that the predictors may have a different point value, or by assigning the same point value (the same 'weight') to each predictor. Such simplification is known to lead to some loss of predictive accuracy. (73)

1.4.2 Stage 2 –External validation of clinical prediction rules

The validation phase determines the performance of the prediction model or rule in individuals different to those from which the prediction model was derived. The process of external validation involves measuring predictor and outcome values in data from new individuals suspected of having the condition of interest, applying the prediction model to these data and quantifying the models predictive performance. (62)

Three types of external validation are recognised, namely, temporal, geographical and domain validation, each providing a progressively more thorough and independent validation. In temporal validation studies, new individuals may be from the same institution in a different usually later, time period. Geographical validation examines the transportability or generalizability of the predictive performance of the prediction model to other institutes or countries and domain validation to very different individuals from those for whom it is developed (i.e. validating a prediction model derived in secondary care to individuals in a primary care setting). Recently methods for quantifying the degree of relatedness between derivation and validation samples have been developed, facilitating the interpretation of prediction model transportability to other settings. (74)

Key aspects of predictive performance to be determined in validation studies include discrimination and calibration. (63) Calibration reflects the agreement between predictions from the rule and observed outcomes. Generally, a rule is said to be well calibrated if, for every group of 100 individuals, each with a predicted risk of $x\%$, close to x have the outcome of interest. Commonly used methods to assess calibration include calibration in the large, the calibration slope and the Hosmer-Lemeshow goodness of fit test. Calibration in the large compares the mean of all predicted risks with the mean observed risks and indicates the extent that predictions are systematically too low or too high. (63) The Hosmer-Lemeshow test examines how well the percentage of observed events matches the percentage of predicted events over deciles of predicted risk. Limitations of this test have been well described and include its dependence on arbitrary groupings of individuals and poor power in small datasets. Furthermore, by only providing a p value, the test conveys no indication of the magnitude or direction of any miscalibration. Consequently, a calibration plot is the preferred methods for assessing calibration. (1, 75) Calibration plots categorise individuals into groups according to predicted risk. The observed risk (proportion of individuals with an event) is calculated for each group, and the predicted risk (the average risk score for individuals in each group) is plotted against the observed risk for each group. The plot displays the direction and magnitude of miscalibration across the probability range.

Discrimination refers to the ability of a prediction rule to differentiate between those who do or do not experience the outcome event. A model has perfect discrimination if the predicted risks for all individuals who have the outcome are higher than those for individuals who do not experience the outcome. Though many indices of discrimination exist, the discriminative ability of a prediction rule is commonly estimated by the concordance index (c-statistic) which is a measure of concordance between prediction rule based risk estimates and observed events. (76) The c-statistic is identical to the area under the receiver operating characteristic curve (AUROC) for logistic regression models. Simplistically, the AUROC represents how likely it is that the test will rank two individuals, one with the event and one without the event, in the correct order across all possible thresholds. (77) While the AUROC summarises the discrimination of a prediction rule in a single number, it lacks clinical interpretability because it does not effectively balance misclassification errors. By measuring performance over all thresholds it includes both those clinically logical and clinically illogical. In addition to measures of discrimination and calibration, measures of overall performance, including the Brier score or R^2 (a measure of variation explained by the model), may be reported. (1)

For diagnostic rules, measures of classification performance such as sensitivity, specificity and predictive values are often presented. Though clinically useful measures, predictive values are highly dependent on the prevalence of the condition in the population. Sensitivity and specificity are generally believed to be unaffected by disease prevalence, though this has recently been shown not to be the case. (78) The use of these measures of performance relies on identification of a cut-off or test threshold to classify individuals as low or high risk. Defining the threshold is based on clinical judgment as to the consequences of false positive decisions (such as unnecessary testing or possible over diagnosis and treatment) and false negative decisions (possible under diagnosis and under treatment), with the optimal threshold defined by the balance between the harm of a false positive classification and the benefit of a true positive classification. If a false positive decision is of relatively low importance in comparison to a false negative decision, a low threshold should be used. For example, unnecessary testing for a patient with chest pain without acute coronary syndrome (false positive) should be avoided but has less clinical significance than withholding treatment in those with disease (false negative). The decision threshold of 2% reflects the relative weight of these errors. (79) However, it may be difficult to define an optimal threshold since empirical evidence for the relative weights of benefits and harms is often not available. (63) Recently, a novel approach (based on clinical vignettes) to determining clinically sensible thresholds for clinical prediction rules has been proposed. (80)

Recently, several new measures aiming to capture the clinical consequences of a particular level of discrimination or degree of miscalibration have been proposed to evaluate the performance of prediction rules. Such approaches include decision curve analysis and relative utility. In decision curve analysis, a single probability threshold is used to weight false negative and false positive classifications. The performance of a model at a single threshold can then be summarised as a weighted sum of true minus false positive classifications (net benefit). (81) For comparison of different models or to quantify the clinical benefit of adding a new predictor to an existing model, methods based on the concept of reclassification of individuals across predefined risk categories have been developed. The net reclassification improvement (NRI) is a commonly used measure of reclassification, (82) but has been shown to be highly sensitive to the selection of thresholds used to define risk categories (83) and to have other limitations. (82) Consequently measures which appropriately weight false positive and negative decisions are preferred. (1) These include three mathematically interconverted measures, namely, the weighted net reclassification improvement, change in net benefit and change in relative utility.

These measures use a harm-benefit ratio to define the weights of true positive and false positive classifications to calculate a single summary measure of reclassification. (84)

Studies validating a clinical prediction rule may be comparative, evaluating both the prediction rule of interest and alternate rules or diagnostic pathways in the same study population against a common reference standard. (1) In the most efficient design of so called 'direct' or 'head to head' comparative studies, the alternate 'tests' and reference standard are applied to all study participants, (85) but randomised designs in which all patients undergo the reference standard and are then randomly assigned to only one of the alternate tests, may be utilised when the tests being compared interfere with each other. In direct comparative accuracy studies, the study design and analytical strategy depend upon the intended role of the prediction rule – that is, how the prediction rule will be positioned to alter the existing diagnostic pathway. (86) A CPR may assume the following roles in a diagnostic pathway: replacement of the existing strategy; add-on; or triage. (87) As an add-on test the prediction rule is used concomitantly with or after the existing testing strategy with the aim of correctly identifying patients with a false negative or positive test on the existing testing strategy. Alternatively, a prediction rule may assume a triage role where it is used before an existing test with the results of the prediction rule determining which patients will then undergo the existing test. As a triage test, the prediction rule may not be intended to improve the diagnostic accuracy of the existing testing strategy, but rather to reduce unnecessary further diagnostic testing. In early evaluations of prediction rules, the role may not be clear, but as the prediction rule progresses into validation studies and the performance of the prediction rule is better understood, the role of the prediction rule becomes apparent. The role may vary according to the context in which it is to be used (e.g. an add-on test in the primary care setting or a triage test in the emergency department), and it may change over time. Identification of the role of the prediction rule is essential for ensuring validation studies are efficiently designed and the most appropriate measure of accuracy and statistical method of analysis are used. (86)

In addition to assessment of the accuracy of diagnostic prediction rules, quantification of their added or incremental value over and above clinical judgment alone may also be of interest. Studies of the incremental value of diagnostic tests aim to discern what the test adds to the diagnostic process over and above standardly available diagnostic information. (32) The incremental value will depend on how much information is already available from the diagnostic workup and it may be that the diagnostic information provided by a particular prediction rule is already conveyed by the previous test results.

Evaluation of the added value of a prediction rule over and above clinical judgment involves comparing multivariable models of clinical judgment with and without the clinical prediction rule as a covariate. An example of such a study is that of Broekhuizen et al., (12) who determined the added value of a prediction model for the diagnosis of chronic obstructive pulmonary disease (COPD) over clinical judgment. Using data from a diagnostic cohort study of patients with possible COPD, physicians' estimates of the probability of COPD obtained after initial examination and history taking were entered into a multivariable logistic regression analysis with COPD as the diagnostic outcome (model 1). A second model was then constructed including the physicians estimate and the estimate of the probability from the prediction model. To quantify the added discriminative value of the prediction model over the clinician, the area under the ROC (AUROC) curve of the two models was compared. Though the difference between the AUROC for the two models is the most familiar statistic for estimating incremental value, and there is broad agreement about its usefulness as a descriptive measure, (88) it has been criticised as being insensitive to detecting small improvements in model performance when a new test is added to a model that already includes important predictors. (77) Subsequently, the use of utility measures (decision curve analysis) and reclassification metrics, such as those discussed above, has been recommended for assessing added value. (1)

1.4.3 Stage 3 – Assessment of the impact of clinical prediction rules

In this stage of development, the extent to which a clinical prediction model or rule is used and affects decision making or health outcomes is quantified. (5, 64) To assess the impact of a clinical prediction model or rule a comparative study is necessary. Study outcomes are measured in a control group exposed to care or management without the use of information from a prediction model or rule and an experimental group exposed to care provided with a prediction model or rule. In these studies an assistive or directive approach may be taken. In the assistive approach, estimated probabilities of the target condition are provided without recommending a course of action. In the directive approach, an explicit management recommendation is provided (with or without the underlying probability estimate). While it is generally felt that assistive approaches are more respecting of clinicians' autonomy, and therefore the information provided by the model more likely to be considered by clinicians, the directive approach is believed to have greater potential to influence clinicians' behaviour (5, 89) and is often preferred by clinicians themselves. (90) However, head to head comparisons between these two approaches are lacking.

Empirical studies of the impact of diagnostic prediction rules may utilise randomised or non-randomised designs, though randomised trials are preferred. (5) The comparison between intervention groups is scientifically strongest when a cluster randomised controlled trial is used, as randomising clusters of clinicians or centres avoids bias owing to learning effects that may result when individual clinicians or patients are randomised. (66) Given the significant challenges and practicalities of conducting cluster randomised trials, alternative randomised designs may be used, including the stepped-wedge cluster randomised trial which compares individuals outcomes between clusters which first apply care as usual and subsequently, at randomly allocated time points, management based on the clinical prediction rule. (10) Often, non-randomised study designs are used, including before-after studies which compare individuals' outcomes in those treated conventionally in an earlier period and those treated in a later period, after introduction of the prediction rule. When the effect of a diagnostic prediction rule on clinicians' behaviour or decision making is of key interest, cross sectional studies may suffice. In this approach, clinicians can be randomised either to exposure to or no exposure to a diagnostic prediction rule and their diagnostic or management decisions are compared. Qualitative studies may also be used, alone or as an adjunct to quantitative studies to describe the interplay between the prediction rule, the context in which it is used and the outcome.

The impact of diagnostic prediction rules may also be estimated indirectly using decision analytic modelling. These models extrapolate the link between the predicted accuracy of the prediction model (and associated uncertainty) and clinical management and downstream patient outcomes using data about the effectiveness of treatments from randomised therapeutic trials or meta-analyses. (91) The use of models to evaluate the effect of diagnostic prediction rules offers several advantages over randomised trials, namely, they are relatively quick to perform and lower cost. It has been posited that models could be used as a preliminary step in the evaluation of diagnostic tests, with results guiding further randomised trial evaluation. (92, 93) The main limitation of this approach is that such models only provide indirect evidence of the effects that diagnostic prediction rules may have on patient health. Further, validity of the model is limited by the availability and quality of existing evidence with probability estimates from observational studies potentially subject to bias and confounding, and the need to extrapolate the results of several studies.

1.5 Potential harms and unintended consequences of clinical prediction rules

Potential harms and unintended consequences of clinical prediction rules are infrequently mentioned in the literature but may arise directly from their use as a strategy to facilitate

selective testing or improve accuracy, or via the unintended consequences of their use on clinicians' reasoning processes or clinician patient relations. Diagnostic prediction rules are often developed as tools to decrease unneeded testing in situations where diagnostic tests are overused. While such prediction rules aim to reduce testing only in those who do not have the disease of interest, individuals who do have disease may be missed and potentially exposed to the harms of delayed or missed diagnosis. For example, in a study of the Wells pulmonary embolism score, application of the score increased the proportion of study participants classified as not having disease (and therefore not requiring further testing) compared to clinical judgment, however, this potential reduction in testing was offset by missing a greater proportion of cases of disease. (94) It has also been suggested that use of prediction rules may, over time, diminish skills in clinical diagnostic reasoning and promote 'intellectual laziness', particularly among trainee clinicians where CPRS may be used as teaching aids. (95) However, this anecdotal claim has not, to my knowledge, been the subject of study. Further, concerns about how diagnostic decision support will affect malpractice litigation have been raised. (96) Research to address this has found that use of a decision aid by a clinician did not affect judgments of malpractice in a mock jury trial, but use did affect ratings of punitiveness. (97) Clinicians' use of clinical prediction rules may also affect the clinician-patient relationship. It has been shown that clinicians using a computerised decision support system are perceived as less capable than clinicians using unaided judgment. (98) How this may affect patient behaviour (e.g. treatment compliance) and health outcomes, however, is unclear.

1.6 Standards of evidence for clinical prediction rules

Though evidence of successful validation and, preferably, of clinical impact should exist before a diagnostic prediction rule is considered for adoption into practice, few derived clinical prediction rules ever undergo external validation, and fewer still impact assessment. The pattern of limited validation and even more limited impact analysis has been demonstrated in several analyses of prediction rule research within and across clinical settings. (45, 99, 100) In a series of reviews conducted over three time periods since 1981, the stages of development of prognostic models published in six leading general medical journals was examined. (5, 40, 45, 101) In the latest review period (from 2006 to 2009), the vast majority of prediction rule studies identified, described the derivation of a prediction model. (61 of 84 studies) Only one quarter (21 of 84 studies) reported an external validation of a prognostic model and only two of the eighty-four studies assessed the impact of use of a prognostic model. This situation does not appear to have changed over time, with the reviews using the same methods performed in earlier time periods reporting similar results. In a descriptive analysis of clinical prediction rules

(n=434) relevant to primary care identified for inclusion in a recently established register, just over half (55%) of the derived rules were evaluated in at least one validation study, but only 12 rules (2.8%) were evaluated in an impact study. (44) This study also found differences in the stages of development of the prediction rules across clinical domains with some clinical areas being associated with more derivation than validation studies and vice versa, though assessment of prediction rule impact was generally rare or non-existent. (44) The reasons for this have not been explored, but most likely relate to the relative ease with which prediction models can be derived (102) and the progressively more difficult and resource consuming research effort required to assess validity and impact. Some have argued that many prediction rules are derived simply for the sake of publication, with developers having little regard for clinical need or intention to further evaluate or implement. (102, 103)

1.7 Methodological standards of published clinical prediction rules

Despite a substantial body of methodological literature and published guidance on how to perform and report prediction research, (1, 2, 5, 10, 42, 62, 64, 101) numerous reviews appraising the methodological conduct and quality of reporting of prediction rule studies have consistently found the methodology and reporting of such studies to be suboptimal. (104-107) In the most recent review of 78 published studies describing the external validation of clinical prediction rules, the quality of reporting was judged to be very poor with details needed to objectively judge the quality of the study either not reported, or inadequately reported. (105) Furthermore, the majority of studies were characterised by poor design, inappropriate handling and acknowledgement of missing data and absence of information on model calibration. (105) Another review assessed the reporting and conduct of prediction studies published in six high impact general medical journals in 2008. (104) The authors of this review concluded that the majority of prediction studies do not follow current methodological recommendations for clinical prediction research, and that improvements in reporting and conduct are clearly needed. Other reviews of prediction models in specific clinical conditions similarly conclude that the methods and reporting of prediction research is 'worryingly poor' and likely to limit the reliability and applicability of such research. (106, 107)

1.8 The use of clinical prediction rules

Despite the abundance of clinical prediction rules in the published literature, it is often stated that few are widely implemented or used in clinical practice. (5, 99, 102, 108) While this may be appropriate for the majority of diagnostic prediction rules for which evidence of successful validation, let alone impact on patient health, is lacking, it may also be the case for tools for which validity and positive clinical benefit has been demonstrated. (109, 110)

1.8.1 Frequency of use

Only a small number of published studies have assessed the extent to which diagnostic prediction rules are used by clinicians in practice. Using survey methods, these studies have reported: a) the frequency with which prediction rules are used as a diagnostic strategy in comparison to other strategies that may be employed at the same stage of diagnostic reasoning; or b) clinicians' self-reported use of a specific diagnostic prediction rule with which they are familiar. From these studies the level of appropriate use, that is, use in instances where a prediction rule exists and it is appropriate to apply it, is unclear. In the first of these studies, investigators analysed data collected by six general practitioners after 300 consultations and found that clinical prediction rules were the most infrequently used strategy employed by clinicians during the refinement stage of diagnosis. (111) However, it is not clear whether clinicians' low preference for prediction rules is because there are no prediction rules available for the conditions encountered or, in some cases, that application of available prediction rules is inappropriate. In a series of surveys of emergency physicians regarding their use of the Ottawa Ankle Rules, the Canadian Cervical-spine Rule and the Canadian Computed Tomography Head Rule (CCHR), self-reported frequent use among physicians in Canada (where the prediction rules were developed) was high (ranging from 57% for the CCHR, to 90% for the OAR. (96, 109, 112-114) However use of these rules among clinicians from other countries was much lower. (90, 112) In a survey of 401 UK general practitioners, self-reported frequency of use of any prediction rules (either prognostic or diagnostic) varied widely by clinical domain and by particular prediction rules within a clinical domain. (46) For instance, most respondents reported using prediction rules for cardiovascular disease and depression but prediction rules for cancer were infrequently utilised. In a survey of emergency physicians, only half of all respondents reported using a prediction rule to estimate the pre-test probability of pulmonary embolism in more than half of applicable cases. (115) Low response rates in the survey studies and self-reported use of clinical prediction rules (which do not necessarily accurately reflect actual behaviour) make interpretation of the findings of these studies difficult.

1.8.2 Barriers and facilitators to use of clinical prediction rules

Numerous barriers to the use of clinical prediction rules have been identified. Barriers are reported either: 1) in survey studies of clinicians, some of which perform multivariate statistical analysis to identify barriers or facilitators that make the largest contribution to clinicians intention to implement prediction rules into practice, and 2) in opinion and editorial papers which often suggest barriers to use of clinical prediction rules based on research conducted with other evidence innovations such as clinical practice guidelines and shared decision making. Using a framework for considering barriers to evidence uptake, these barriers

may be classified into the domains of knowledge, attitude and behaviour. (116, 117) Table 1.1 provides an overview of the literature informed potential barriers to the adoption of clinical prediction rules using this schema. Recognition of the facilitators and barriers to implementation of clinical prediction rules in practice is necessary to improve understanding on how to effectively translate clinical prediction rules into clinical practice.

Opposing knowledge acquisition are clinicians' lack of awareness, multiple rules for the same or similar outcome, and unfamiliarity with a prediction rule. (46, 118) Lack of awareness may be due, paradoxically, to the overwhelming volume of prediction research available and to suboptimal efforts to disseminate prediction rule research. With few exceptions, (90) the majority of clinical prediction rule studies are disseminated through passive methods such as journal publication or presentation at scientific meetings which are generally reported to be minimally effective strategy for altering clinical behaviour. (119, 120) Systematic reviews of studies developing prediction rules have identified numerous models for predicting the same outcome or target population, (121-123) making it difficult to decide which one to use. (103, 124, 125) In multivariate analysis, younger age, full time employment and employment in a teaching hospital have been found to be significant predictors of awareness of prediction rules for cervical spine and head injury. (96, 112, 126)

Barriers that prevent change in attitude identified in survey studies include lack of confidence in the prediction rule (115, 118) and conviction that a clinician's own judgment is superior to a prediction rule. (46, 115, 118, 127) The perception that prediction rules oversimplify the assessment process has been reported to be a barrier to use in some studies (128) but not others. (114) Attitudes that may impede the use of prediction rules has been found to vary across countries, with clinicians in the United States, for instance, indicating that the use of prediction rules protects against malpractice lawsuits far less often than clinicians from other countries. (114) In the only available survey of non-medical clinicians, inhibitive attitudes towards prediction rules for a particular condition included views that available prediction rules were not sufficiently developed, were rarely generalizable, oversimplified the reasoning process or were not needed by experienced clinicians. (95)

Behaviour change may be impeded by external pressures that favour the inertia of the status-quo, including environmental, patient and institutional factors, and features of the prediction rule itself. In survey studies, clinicians' report agreement with use of a prediction rule when a prediction rule helps to save time and is easy to use and remember. (95, 109, 126) Ratings that a rule is an efficient use of time and not too much trouble to apply were also significant

predictors of use of a prediction rule for cervical spine injury in a study of emergency physicians. (96) Perceived reduction in patient satisfaction has also been reported by clinicians as a barrier to use. (118, 129) While environmental and institutional factors (e.g. lack of time or institutional support) have been shown to be barriers to the use of clinical practice guidelines, (116) and have been suggested as barriers to use of clinical prediction rules specifically, they have been rarely studied or reported. (129)

The features of the clinical prediction rule itself that may facilitate or impede adoption or use in practice are unclear. Methodological papers, editorials and survey studies have suggested that rule presentation, face validity and complexity of the clinical prediction rule may influence use, but empirical studies linking these features to use (or non-use) are uncommon. (1, 45, 130, 131) For a clinical prediction rule to have face validity it should include predictors that are anticipated by clinicians (often termed content validity) and are biologically plausible. There should be no obvious omissions, and the way the predictor variables are organised should seem appropriate for the purpose of the rule. (130, 131) While absence of predictors that clinicians feel are important may contribute to non-use of a specific clinical prediction rules (118, 128), inclusion of predictors that don't have a logical relationship with the dependent variable also presents as a threat to acceptance and implementation. (95)

While 'ease of use' is frequently mentioned as necessary for clinical prediction rule adoption into practice (45, 109, 118, 130, 131), what makes a clinical prediction rule 'easy to use' is not clear and is predominantly a subjective judgment by the user. The format used to present the prediction model to users may be a factor in perceptions of useability, particularly for clinical prediction rules that are not computerised. (131) Inextricably linked to the presentation format of clinical prediction rules is their complexity, which may include factors such as the number of predictors included in the clinical prediction rule, the ease with which information on the predictors can be obtained and the calculations required to obtain an output from the clinical prediction rule. Scoring systems are a frequently used presentation format of diagnostic prediction rules that, when not available electronically, require users to sum variably scored predictors, or more simply to count the number of predictors. Uncomplicated calculations or simple graphical aids are likely to facilitate use in practice (95, 132), and may prevent calculation errors when deriving the output of the clinical prediction rule. The number of predictors included in a clinical prediction rule is another modifiable feature that may enhance the perceived usefulness of a clinical prediction rule. In a survey of physiotherapists, clinical prediction rules with a 'large' number of predictors were viewed negatively as were the

inclusion of predictor variables that are difficult to obtain in a timely fashion without sophisticated equipment. (132)

1.8.3 How diagnostic prediction rules may be used during clinical encounters

Little is known about how clinicians actually use diagnostic prediction rules in practice. Limited research among physiotherapists and emergency physicians suggests that a prediction rule may be used as a 'second opinion' or 'safety net' to validate clinicians' decision-making. (95, 133, 134) Among clinicians intent upon using a particular prediction rule, research suggests that the prediction rule may be variably implemented. (109) For example, a prediction rule may be used in part, meaning that only some predictors included in the prediction rule are considered and data collected. This may be intentional (e.g. if predictor data is not available or difficult to obtain) or accidental (e.g. if the clinician is applying the clinical prediction rule by memory but is unable to recall all the included predictors). (109, 115) Further, a clinician may not use the prediction rule as the primary determinant of decisions, incorporating other clinical features known to be unrelated to the outcome (e.g. cracking sound for ankle fracture) or correlated with the outcome but of no added value over and above the variables in the rule (e.g. age for ankle fracture) and subjectively adjusting the rule for these factors. (109, 135) A clinical prediction rule may be applied inconsistently, being applied only in some clinical situations for which it would be appropriate and not others. Or it may be applied incorrectly. For instance, a diagnostic prediction rule may be proposed as a '1-way rule' designed to rule out a particular condition. If the rule is negative, no x-ray should be performed. If the prediction rule is positive, however, it means only that the condition cannot be confidently ruled out, not that an x-ray should be performed. Further, the prediction rule output may be derived incorrectly, for instance by incorrectly adding a scoring system. Where clinicians in practice deviate from the application of a prediction rule as specified by rule developers, it is likely to alter the predictive value and performance of the rule in accuracy studies.

Table 1.1. Literature informed potential barriers to the adoption of clinical prediction rules in practice

Theme	Subtheme	Barrier
Knowledge	Awareness	Clinician unaware of the existence of CPRs for certain conditions Clinician unable to select CPR when multiple CPRs for a condition are available
	Familiarity	Clinician not familiar enough with a CPR to implement it
Attitudes	Agreement with CPRs in general or a specific CPR	Clinician perception that CPRs threaten autonomy Clinician perception that CPRs oversimplify the clinical assessment process Clinician conviction that clinical judgment is superior to CPRs Clinician belief that clinical judgment is not error prone and therefore does not need to be 'fixed' Clinician belief that use of CPRs leads to intellectual laziness Clinician belief that CPRs not sufficiently developed Clinician belief that patient will find them less capable if using a CPR Clinician belief that CPRs are only relevant for inexperienced clinicians Clinician belief that probability estimates are not helpful for decision making Clinician dislike of the term 'rule'
	Outcome expectancy	Belief that use of CPRs will not lead to improved patient or process outcomes Fear of unintended consequences of use Belief that the information from the CPR is not sufficient to change clinical decisions Uncertainty about the effects of use in patients with atypical presentation
	Self-efficacy	Clinician belief that CPR is too difficult to use Clinician uncertainty as to how to interpret or use CPR output
	Motivation	Lack of motivation to use CPR
Behaviour	Patient factors	Patients expectations are not consistent with the CPR
	Features of the CPR	Perception that CPR too complicated or complex Perception that CPR not an efficient use of time Perception that CPR does not have face validity – predictors clinicians consider important are missing Perception that CPR is too much trouble to apply CPR requires input of difficult to obtain predictor data Perception that use of CPR does not fit in with normal work flow Perception that CPR not generalizable to their patient Perception that CPR is static and does not take into account the dynamic nature of clinical practice
	Environmental factors	Lack of time Lack of organisational support Lack of peer support for use Perceived increased risk of litigation with use

1.9 Gaps in the evidence base and research justification

Diagnostic prediction rules are tools developed to support clinical judgment during the diagnostic workup. They are purported to have the potential to improve clinical decision making and ultimately patient outcomes or to provide other benefits (such as reduced resource use) without compromising patient wellbeing. Studies describing the derivation of clinical prediction rules are abundant in the medical literature, and health care providers and policy makers are increasingly recommending their use within clinical practice guidelines. However, the value of diagnostic clinical prediction rules is unclear.

A key step in the evaluation of a diagnostic prediction rule is investigation of its diagnostic performance relative to a reference standard. Ideally these studies also compare the diagnostic performance of the diagnostic prediction rule to the performance of the existing test or testing pathways. (85, 87) In the diagnostic setting this is most likely to be the judgment of the clinician. Comparing diagnostic performance and other features of the diagnostic prediction rule when it is applied independently of clinical judgment, and clinical judgment alone, in the early phase of development can assist in defining how the prediction rule may be used (as a replacement, add-on or triage test) and in guiding further evaluation. Though at the commencement of this thesis, the comparative accuracy of probability estimates or diagnosis provided by clinical judgment and those provided by statistical models had been extensively studied across diverse fields, the findings of this large body of research were difficult to apply to clinical practice. Primarily, the metrics used in these comparative studies to judge the superiority of the alternate methods precluded consideration of the clinical importance of any differences detected, for example, the occurrence and relative importance of false positive and false negative findings. Given the limitations of the existing research, a systematic review of the comparative performance of diagnostic prediction rules and clinical judgment, and the added value of diagnostic prediction rules was considered warranted.

Ideally, diagnostic prediction rules should only be introduced into practice if evidence indicates that they are more likely than the existing diagnostic process to improve patient health or to maintain patient health while offering other benefits. (62) The focus of prediction rule research to date, however, has been to establish the accuracy of prediction rules compared to the prevailing reference standard. Often this information is used as the basis for judgments on the clinical value of a prediction rule and/or to inform recommendations for its use in practice. Given the numerous mechanisms by which prediction rules may act to alter patient health beyond those arising from superior accuracy, reliance on evidence of accuracy (which correlates poorly with patient outcomes) to inform of the clinical value of diagnostic prediction

rules is problematic. Quantification of whether, and the extent to which, the use of a diagnostic prediction rule, either as an add-on test, a replacement for an existing testing strategy, or as a triage test to determine who undergoes the existing test, alters health outcomes is necessary and, to this end, a systematic review of studies comparing care provided with and without a diagnostic prediction rule was considered justified.

Diagnostic prediction rules have been proposed as a strategy for assisting clinicians in one clinical area where diagnosis is especially challenging, namely differentiating a child with a serious bacterial infection from one with a self-limiting illness. This is particularly difficult in the primary care setting where the incidence rates of serious infection are low and children usually present in the early stage of illness where signs and symptoms of serious and non-serious infections appear similar. Though, in the early stages of this thesis, several clinical prediction rules for the identification of children with serious infection had been developed, these tools were of questionable applicability to the primary care setting (for example, they included tests not generally available) and/or had key methodological limitations. Subsequently clinicians across a range of developed countries had identified the management of children with fever and possible serious infection as the clinical problem they would most like to approach with a well-designed prediction tool. (136) To address this need, the derivation of a prediction model based on features from history and physical examination, with external validation was planned. At the same time, C-reactive protein, an inflammatory marker was being promoted as a useful test for the detection of serious bacterial infection in children. Point of care versions of the test were available and being used or promoted in many parts of the world. However, the diagnostic accuracy of the marker had not been determined. It was my view, therefore, that a review of the accuracy of C-reactive protein was justified, with a subsequent study to determine the added value of C-reactive protein over clinical characteristics readily available in the primary care setting of key interest.

Few clinical prediction rules are used in clinical practice. While many barriers to the implementation of prediction rules relating to the clinician or the environment in which they operate have been identified, features of the prediction rule itself are likely to be important in facilitating or impeding uptake. Intuitively, prediction rules that are easier to use are more likely to be adopted into practice. Given the potential for simplification to facilitate the uptake of prediction rules, an examination of the effect of several methods of simplification of a diagnostic prediction rule on accuracy and risk classification was believed to be warranted.

1.10 Research aims and thesis overview

1.10.1 Research aims

This thesis addresses a common challenge facing clinicians, namely, how to efficiently and accurately arrive at a diagnosis or clinical decision. The primary goals of this thesis were to determine the clinical value of diagnostic prediction rules developed to assist clinicians in a range of clinical scenarios and to assist primary care clinicians in their diagnostic management of children with possible serious bacterial infection. Within these goals, my aims were;

Aim 1: To determine the comparative diagnostic performance of diagnostic prediction rules (when applied to patient data independently of clinicians' judgment) and clinical judgment versus a reference standard and the added value of diagnostic prediction rules beyond clinical assessment.

Aim 2: To determine the effect of care provided when a clinician has access to a diagnostic prediction rule compared to care provided without a diagnostic prediction rule on patient and process outcomes.

Aim 3: To assist primary care clinicians in the differentiation of children with serious bacterial infections from children without serious infection. This is to be achieved through the derivation and validation of a clinical prediction rule for the identification of serious bacterial infections generally and pneumonia specifically, in children presenting to primary care, and through the determination of the diagnostic accuracy, independent and added value of the inflammatory biomarker C-reactive protein. An existing dataset of children presenting to a paediatric assessment unit in the United Kingdom will be used for the derivation study and study of the added value of C-reactive protein.

Aim 4: To assess the effect of prediction rule simplification on performance.

1.10.2 Thesis outline

The proposed research aims are addressed in the following five chapters. Three chapters have been published in peer reviewed journals (Chapters 2, 5 and 6) and one has been submitted and is under external review (Chapter 3). In order to make the published chapters easier to read they have been formatted in a style consistent with the body of the thesis.

Chapter 2 is a systematic review comparing the diagnostic performance of diagnostic prediction rules and clinical judgment with a reference standard.

Chapter 3 presents a systematic review comparing the effects of care provided by clinicians with access to a diagnostic prediction rule, versus care provided without a diagnostic model or rule, on patient and process outcomes.

From this point the thesis focuses on aspects of clinical prediction rule development and application in a clinical area where diagnosis is known to be especially challenging, namely, the identification of serious bacterial infection in children with fever. The nature of this difficult diagnostic problem is introduced in **Chapter 4**.

Four projects were proposed to address the thesis goal to assist primary care clinicians in their diagnostic management of children with possible serious bacterial infection. The first of these was the derivation and validation of a diagnostic prediction rule for the identification of serious infection in children presenting to primary care using an existing dataset. The second and third were projects to determine the accuracy, independent and added value of the inflammatory marker C-reactive protein in the primary care setting using systematic review methodology and by performance of a modelling study using the derived prediction model as the base model for the analysis. The fourth and final project was to test the effects of methods of simplification of the derived model on accuracy and reclassification of individuals across predefined risk categories.

However, the first project to derive and validate a clinical prediction rule to assist primary care clinicians in the diagnostic management of children with possible serious infection could not be completed. During preliminary work to develop the prediction rule, I judged that due to the volume and likely nature of missing data and limitations of methods of dealing with missing data at the time, development of a credible prediction rule was not achievable. Consequently, the study to determine the added value of C-reactive protein beyond information obtained from the history and physical examination could also not be undertaken. In **Chapter 4**, the rationale for the proposed but unexecuted studies is briefly discussed, as are the reasons for this eventuality. The systematic review to determine the accuracy and independent value of C-reactive protein was completed and the project to test the effects of simplification was performed with data obtained from an alternate source. The remainder of the thesis is therefore comprised of the following chapters.

Chapter 5 presents the results of the systematic review of the accuracy and independent value of C-reactive protein for the identification of serious bacterial and bacterial infection in non hospitalised children.

Chapter 6 of this thesis presents the study undertaken to ascertain the effects of various methods of simplification of a diagnostic prediction rule on performance. As the prediction rule for serious infection in children could not be derived and used for this study, and efforts to obtain data in the clinical area of interest were unsuccessful, the effect of simplification was evaluated in an existing dataset from which a prediction rule had been developed for the risk stratification of individuals with chest pain and suspected cardiovascular event.

The concluding chapter of this thesis, **Chapter 7** summarises the main findings and discusses limitations and implications of the research conducted.

Chapter 2 The comparative performance of diagnostic prediction rules and clinical judgment

This chapter presents an article published in PLOS One on 3rd June 2015.

Sanders S, Doust, J, Glasziou P. A systematic review of studies comparing diagnostic clinical prediction rules with clinical judgment. PLOS One. 2015;10(6):e0128233.

Doi:10.1371/journal.pone.0128233.

© 2015 Sanders et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Minor modifications to the methods section of the published manuscript have been made in response to comments received during the thesis examination process. Details of the modifications and their location in the manuscript are tabulated in Section 2.6.4 at the end of this chapter.

2.1 Preface to Chapter 2

Quantifying the diagnostic performance of a diagnostic prediction rule is a key step in its evaluation. Comparing the diagnostic performance and other features of the prediction rule with the existing diagnostic test or pathway can assist in defining how the prediction rule may be used and in guiding its further evaluation. The following chapter describes a systematic review which aimed to determine the comparative performance of clinical judgment and diagnostic prediction rules (when the diagnostic prediction rule is applied to the study data independently of clinical judgment). Diagnostic performance is compared in terms of a) the ability of the two methods to classify individuals as not having disease and thereby avoiding further testing, referral or treatment and b) the proportion of those individuals classified as not having the condition of interest that actually do. The comparative discriminative ability of both judgment methods in terms of the traditional paired summary statistics sensitivity and specificity is also presented.

2.2 Abstract

Background: Diagnostic clinical prediction rules (CPRs) are developed to improve diagnosis or decrease diagnostic testing. Whether, and in what situations diagnostic CPRs improve upon clinical judgment is unclear.

Methods and findings: We searched MEDLINE, Embase and CINAHL, with supplementary citation and reference checking for studies comparing CPRs and clinical judgment against a current objective reference standard. We report 1) the proportion of study participants classified as not having disease and hence may avoid further testing and or treatment and, 2) the proportion, among those classified as not having disease, who do (missed diagnoses) by both approaches.

31 studies of 13 medical conditions were included, with 46 comparisons between CPRs and clinical judgment. In 2 comparisons (4%), CPRs reduced the proportion of missed diagnoses, but this was offset by classifying a larger proportion of study participants as having disease (more false positives). In 36 comparisons (78%) the proportion of diagnoses missed by CPRs and clinical judgment was similar, and in 9 of these, the CPRs classified a larger proportion of participants as not having disease (fewer false positives). In 8 comparisons (17%), the proportion of diagnoses missed by the CPRs was greater. This was offset by classifying a smaller proportion of participants as having the disease (fewer false positives) in 2 comparisons. There were no comparisons where the CPR missed a smaller proportion of diagnoses than clinical judgment and classified more participants as not having the disease. The design of the included studies allows evaluation of CPRs when their results are applied independently of clinical judgment. The performance of CPRs, when implemented by clinicians as a support to their judgment may be different.

Conclusions: In the limited studies to date, CPRs are rarely superior to clinical judgment and there is generally a trade-off between the proportion classified as not having disease and the proportion of missed diagnoses. Differences between the two methods of judgment are likely the result of different diagnostic thresholds for positivity. Which is the preferred judgment method for a particular clinical condition depends on the relative benefits and harms of true positive and false positive diagnoses.

2.3 Introduction

Diagnostic clinical prediction rules (CPRs) are tools designed to improve clinical decision making. (2) Theoretically, CPRs, by providing objective estimates of the probability of the presence or absence of disease derived from the statistical analysis of cases with known outcomes and or by suggesting a clinical course of action, can improve the accuracy of diagnosis and or decision making.

Understanding whether and in what situations CPRs improve upon clinical judgment is an important step in the evaluation of CPRs and for the acceptance of CPRs by clinicians. (131) Existing research, which has focused on the comparative performance of CPRs and clinical judgment when both judgment methods are viewed as competing alternatives, is difficult to interpret. One body of research on the relative merits of clinical and statistical prediction has consistently reported the superior accuracy of statistical models over a clinicians ability to integrate the same data and to collect and integrate their preferred data, (36, 55, 58) while another, more recent body of research has found that heuristics – proposed as models of human judgment, are on occasions more accurate than statistical models. (56) It is also difficult to know how to apply the general findings of this research to clinical practice. Many of the reviews of comparative accuracy have summarised findings from diverse professional fields including finance, medicine, psychology and education. Further, judging the clinical utility of clinical judgment and CPRs requires consideration of not just overall accuracy but the consequences of missed diagnoses (false negative) and false positive results. Results of the existing comparative research are generally not reported in a way that allows such evaluation.

We conducted a systematic review of studies that compared the performance of diagnostic CPRs with clinical judgment or the performance of the combination of CPR and clinical judgment versus either alone in the same study participants against a current and objective reference standard.

2.4 Methods

This review was performed following methods detailed in the systematic review protocol and is reported in line with the PRISMA Statement (Appendix A).

2.4.1 Data sources and searches

We searched MEDLINE, Embase and CINAHL from inception to January 2012, with an updated MEDLINE search to March 2013 (Appendix A). No limits were applied to the database searches. We also searched for systematic reviews of diagnostic CPRs using PubMed Clinical Queries. The reference lists of systematic reviews and the included studies were checked. We

conducted forward searches of included studies using Science Citation Index Expanded in Web of Science and checked related citations using PubMed's Related Citations link.

2.4.2 Study selection

We included studies that compared the CPRs with clinical judgment in the same participants using a current and objective reference standard. We also included studies that compared a CPR or clinical judgment alone with the combination of CPR and clinical judgment and modelling studies to determine the added value of CPRs above clinical judgment. The CPR had to have been developed using a method of statistical analysis and tested against clinical judgment in a population different (by time, location or domain) to that from which it was derived. Studies where the CPR and clinical judgment were applied to different individuals (for example, in randomised trials) or were not applied at approximately the same point in the diagnostic pathway were excluded (for example, if the result of a CPR was determined using data collected at first presentation and this was compared to clinical judgment made after further consultation, testing and observation). We excluded studies of CPRs for the diagnosis of disorders across multiple body systems, that are not applied to actual patients, that are used for the interpretation of tests such as ECGs or that are performed in selected samples of patients not consistent with populations for whom use of the CPR is intended.

Titles and abstracts identified by the searches were screened by one reviewer and obviously irrelevant articles excluded. A second reviewer independently screened 15% of the titles and abstracts to ensure that no further studies met the inclusion criteria. After screening, potentially relevant studies were obtained in full text and independently assessed by two reviewers against the review inclusion criteria. Discrepancies were discussed and resolved with a third reviewer.

2.4.3 Data extraction and risk of bias assessment

Two reviewers independently extracted data on the characteristics of the study, the risk of bias and the results using a piloted data collection form. QUADAS-2 (137) was used to assess the risk of bias and concerns regarding applicability in each of the included studies. We added an additional signalling question to identify if clinical judgment and the CPR were determined independently. Discrepancies between reviewers were discussed and resolved by discussion with a third reviewer.

2.4.4 Data synthesis and analysis

We grouped studies where a probability estimate, clinical diagnosis or decision was made by:

a) Clinical judgment alone;

b) Clinical judgment with a method of structured data collection. Clinicians may have collected data on variables contained in the CPR as per the study protocol but calculation of the results of a CPR by the clinician was not anticipated or expected, or occurred after the clinician had provided their probability estimate or diagnosis; or

c) A combination of clinical judgment and clinical prediction rule, where the clinician had access to the results of the CPR but could also use their own judgment or override the CPR.

We also recorded whether the result of the CPR was calculated by the examining clinician or a researcher, the method used to elicit clinical judgment and whether clinical judgment was a clinicians' probability or risk assessment (e.g. low or high risk), a diagnosis or a clinical decision.

Because many clinical prediction rules are developed to either improve the proportion of individuals with a suspected disease classified as not having the disease (thereby decreasing the number of participants undergoing further testing, referral or treatment), or to reduce the number of cases of disease missed by the current diagnostic protocol, the main outcome measures of the review were 1) the percent of study participants classified as not having the disease by the CPR or clinical judgment ($\text{False negative (FN)} + \text{True negative (TN)} / \text{total number of participants in the study (total N)}$). The higher this proportion, the fewer individuals that may undergo further testing, referral and or treatment, and 2) the percent of study participants among those classified by the CPR or clinical judgment as not having the disease who actually have the disease ($\text{FN} / (\text{FN} + \text{TN})$ or 1-negative predictive value). It is desirable that this be as close to 0% as possible (Box 2.1). We also report measures of diagnostic accuracy including the sensitivity ($\text{True positive (TP)} / (\text{TP} + \text{FN})$) and specificity ($\text{True negative (TN)} / (\text{FP} + \text{TN})$) of CPRs and clinical judgment, and present graphically the proportion of all study participants who are classified by CPRs and clinical judgment as having disease who do ($\text{True Positives} / \text{total N}$) and do not ($\text{False Positives} / \text{total N}$) and the proportion of all participants who are classified as not having disease who do ($\text{False Negatives} / \text{total N}$) and do not ($\text{True Negatives} / \text{total N}$). When the output of the CPR or clinical judgment was not a binary decision or action (e.g. the CPR classified individuals as low, moderate or high risk, or clinical judgment was a clinicians' decision to not test or treat, to test, or to test and treat), we dichotomised the output by combining probability estimates that were not 'low' (e.g. moderate and high or possible or probable), and decision or actions that involved tests or treatment (e.g. for a score throat score, the directive output 'culture' and 'culture and treat' were combined and compared to 'no culture or treatment').

We did not perform a meta-analysis due to clinical and statistical heterogeneity. Instead, we synthesised the results of the included studies overall, and by clinical condition (where there were 2 or more studies available) by determining the number of comparisons in which the proportion of participants classified as not having disease and the proportion of missed cases of disease (missed diagnoses) in participants classified as not having disease for CPRs and clinical judgement was similar, greater or lesser. To determine whether there was a difference in the proportion classified as not having disease between CPRs and clinical judgment we conducted a statistical test of the difference between two proportions from dependent samples. To obtain the statistical significance of the relative difference in the proportion classified by CPRs and clinical judgment as not having disease that do, we conducted a test of the strength of association between two proportions (false negative rates) from dependent samples. If studies reported different thresholds for clinical judgment or the CPR, and if the proportions (i.e. the proportion classified as not having disease and the proportion of missed diagnoses) at the different thresholds were both in favour of, or opposed to the CPR or clinical judgment (this only occurred in 1 study included in this review) we reported only the comparison for the threshold with the highest Youden's index ((sensitivity + specificity)-1).

Box 2-1 Outcomes of the review

		Disease		
		Positive	Negative	
Test	Positive	TP	FP	TP + FP
	Negative	FN	TN	FN + TN
		TP + FN	FP + TN	Total N

TP – true positive; FP – false positive; FN – false negative; TN – true negative

The percent of study participants classified as not having the disease by the test (CPR or clinical judgment) = (FN + TN) / Total N

The percent of study participants among those classified by the test (CPR or clinical judgment) as not having the disease who actually have the disease = FN / (FN + TN)

Sensitivity = TP / (TP + FN) Specificity = TN / (FP + TN)

2.5 Results

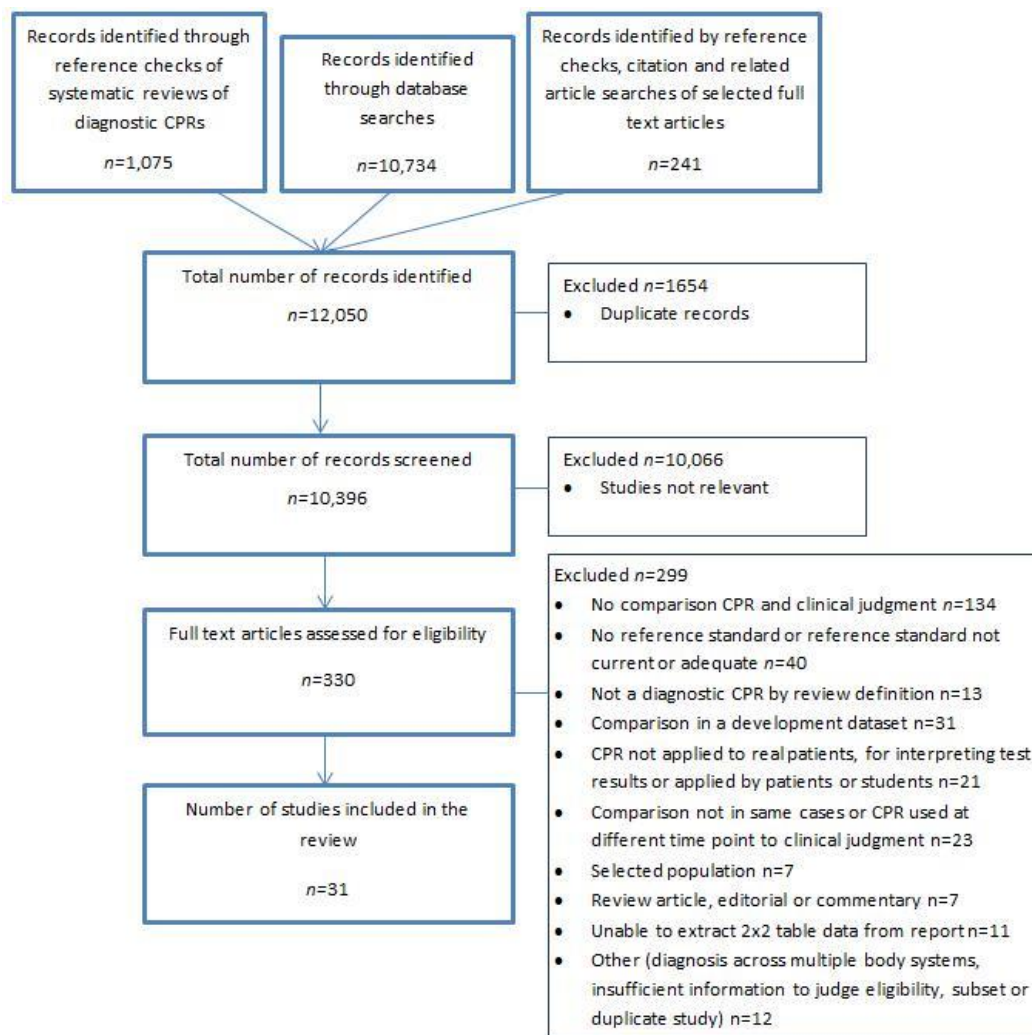
2.5.1 Literature search

Of 10,155 titles and abstracts screened against review eligibility criteria, 330 were obtained in full text and assessed for eligibility by two reviewers. 31 studies (94, 138-167) were included in the review (Figure 2.1).

2.5.2 Study characteristics

The studies addressed a variety of conditions: 9 for pulmonary embolism (PE), (94, 143, 145, 153-155, 160, 162, 163) 6 for deep vein thrombosis (DVT), (140, 141, 146, 151, 158, 167) 3 for streptococcal throat infection, (139, 144, 161) 3 for ankle and/or foot fracture, (138, 152, 164) 2 for acute appendicitis (150, 157) and one each for acute coronary syndrome, (159) pneumonia, (149) head injury in children, (147) cervical spine injury, (166) active pulmonary tuberculosis, (148) malaria, (142) bacteraemia (156) and influenza. (165) (Table 2.1) Twenty five different CPRs were evaluated. The majority (n=16) were derived from logistic regression analysis and the remainder from recursive partitioning analysis (n=3), discriminant analysis (n=2), neural networking (n=1), simple Bayesian analysis (n=2) and an unspecified multivariable analysis (n=1). In just over half of the included studies (n=17), clinical judgment was a clinician's estimate of the probability of the presence of disease or categorisation of a study participant into a risk group (e.g. low, intermediate or high risk). In the remaining studies, clinical judgment was a clinician's diagnosis (n=8), intended management (n=3) or the clinical action taken (n=3). In half of the included studies (n=15) the experience of clinician's estimating the probability of the target disorder or making a diagnosis or management decision was not reported. Ten studies included clinicians with varying levels of experience (e.g. 'post graduates' and 'confirmed emergency physicians'), 3 included specialists only and 3 junior staff only.

Figure 2.1. PRISMA flow diagram of the article selection process



2.5.3 Risk of bias

87% (27/31) of studies were judged to be at high or unclear risk of bias on two or more domains of the QUADAS-2 tool (

Figure 2.2). The most common risk of bias was due to interpretation of the reference standard occurring with knowledge of the index test result. For most studies in which the CPR was applied retrospectively to the data, it was not possible to determine whether researchers were blind to the result of the reference standard test. This is likely to bias results in favour of the CPR. 55% (17/31) of studies were judged to be at high risk of bias on the flow and timing domain. Studies commonly failed to include all enrolled cases in the data analysis or incorporated one of the index tests in the reference standard. Risks of bias assessments for individual studies are shown in Table 2.2.

Table 2.1. Clinical conditions and study comparisons

Clinical condition	Number of studies (number of comparisons)	Methods of estimating a probability, making a diagnosis or management decision being compared (number of comparisons)
Pulmonary embolism (94, 143, 145, 153-155, 160, 162, 163)	9* (16)	CPR versus clinical judgment alone (1) CPR versus clinical judgment + structured data collection (13) CPR versus combination of clinical judgment and CPR (2)
Deep vein thrombosis (140, 141, 146, 151, 158, 167)	6 (7)	CPR versus clinical judgment alone (1) CPR versus clinical judgment + structured data collection (5) CPR versus combination of clinical judgment and CPR (1)
Streptococcal throat infection (139, 144, 161)	3 (5)	CPR versus clinical judgment + structured data collection (5)
Ankle or foot fracture (138, 152, 164)	3 (4)	CPR versus clinical judgment alone (1) CPR versus clinical judgment + structured data collection (3)
Acute appendicitis (150, 157)	2 (2)	CPR versus clinical judgment alone (1) CPR versus combination of clinical judgment and CPR (1)
Acute coronary syndrome (159)	1 (1)	CPR versus clinical judgment + structured data collection (1)
Pneumonia (149)	1 (4)	CPR versus clinical judgment + structured data collection (4)
Abnormalities on computed tomography scan in child with head injury (147)	1 (1)	CPR versus clinical judgment alone (1)
Cervical spine injuries (166)	1 (1)	CPR versus combination of clinical judgment and CPR (1)
Active pulmonary tuberculosis (148)	1 (1)	CPR versus clinical judgment + structured data collection (1)
Malaria (142)	1 (2)	CPR versus clinical judgment alone (2)
Bacteremia (156)	1 (1)	CPR versus clinical judgment + structured data collection (1)
Influenza (165)	1 (1)	CPR versus clinical judgment alone (1)

Figure 2.2. Summary QUADAS-2 risk of bias and applicability judgments

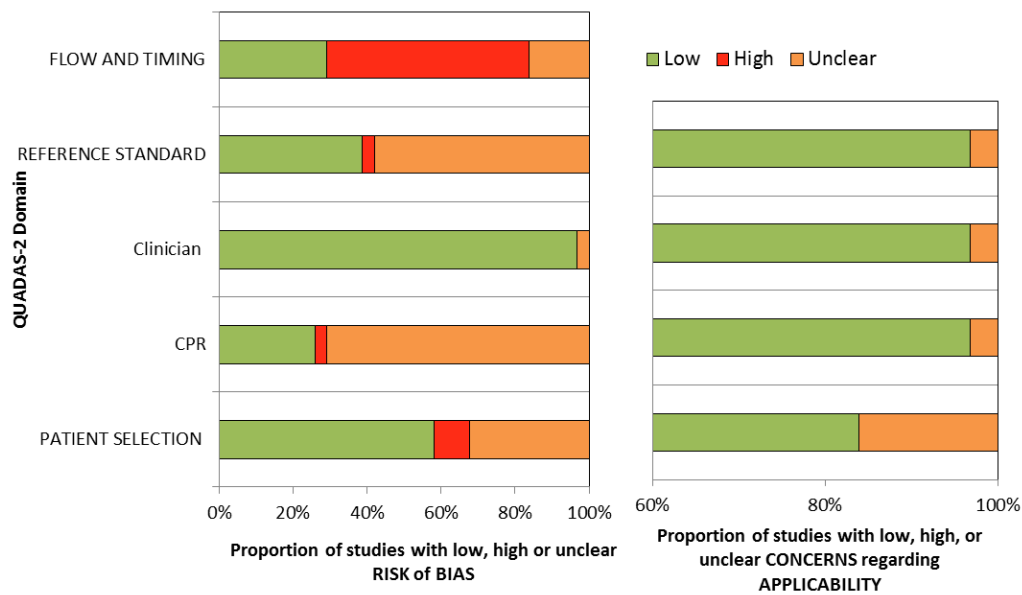


Table 2.2. Risk of bias and applicability concerns for individual studies included in the review

Study	Risk of Bias					Concerns regarding Applicability			
	Patient selection	Index test (CPR) [†]	Index test (clinical judgment)	Reference standard [†]	Flow and timing [‡]	Patient selection	Index test (CPR)	Index test (clinical judgment)	Reference standard
Pulmonary embolism									
Runyon et al., 2005 (162)	Unclear	Unclear	Low	Low	High	Low	Low	Low	Low
Kabrhel et al., 2009 (153)	Low	Unclear	Low	Unclear	High	Unclear	Low	Low	Low
Kline et al., 2008 (155)	Low	Unclear	Low	Unclear	Unclear	Unclear	Low	Low	Low
Kabrhel et al., 2005 (154)	Unclear	Unclear	Low	Unclear	High	Unclear	Low	Low	Low
Carrier et al., 2006 (143)	Low	Unclear	Low	Unclear	High	Low	Low	Low	Low
Chagnon et al., 2002 (145)	Low	Unclear	Low	Unclear	High	Unclear	Low	Low	Low
Penaloza et al., 2012 (160)	Low	Unclear	Low	Unclear	High	Low	Low	Low	Low
Sanson et al., 2000 (163)	Low	Unclear	Low	Unclear	High	Low	Low	Low	Low
Penaloza et al., 2013 (94)	Low	Unclear	Low	Unclear	High	Low	Low	Low	Low
Deep vein thrombosis									
Geersing et al., 2010 (151)	Low	Low	Low	Unclear	High	Low	Low	Low	Low
Bigaroni et al., 2000 (140)	Low	Low	Low	Unclear	Unclear	Low	Low	Low	Low
Miron et al., 2000 (158)	Low	Low	Low	High	High	Low	Low	Low	Low
Blattler et al., 2004 (141)	Low	High	Low	Unclear	Low	Low	Low	Low	Low
Cornuz et al., 2002 (146)	Low	Low	Low	Low	High	Low	Low	Low	Low
Wang et al., 2013 (167)	Unclear	Unclear	Low	Unclear	Unclear	Low	Low	Low	Low

Table 2.2 Continued

Study	Risk of Bias					Concerns regarding Applicability			
	Patient selection	Index test (CPR) [*]	Index test (clinical judgment)	Reference standard [†]	Flow and timing [‡]	Patient selection	Index test (CPR)	Index test (clinical judgment)	Reference standard
Streptococcal throat infection									
Cebul and Poses, 1986 (144)	Unclear	Unclear	Low	Low	High	Low	Low	Low	Low
Rosenberg et al., 2002 (161)	High	Unclear	Low	Low	High	Low	Low	Low	Low
Attia et al., 2001 (139)	Unclear	Unclear	Low	Low	High	Unclear	Low	Low	Low
Ankle or foot fracture									
Glas et al., 2002 (152)	Low	Unclear	Low	Low	Low	Low	Low	Low	Low
Singh-Ranger and Marathias, 1999 (164)	Low	Low	Low	Unclear	Low	Low	Low	Low	Low
Al Omar and Baldwin, 2002 (138)	Unclear	Low	Low	Unclear	Low	Low	Low	Low	Low
Conditions with ≥ 2 studies									
Fenyo, 1987 (150)	Low	Low	Low	Unclear	High	Low	Low	Low	Low
Meltzer et al., 2013 (157)	Unclear	Unclear	Low	Unclear	High	Low	Low	Low	Low
Mitchell et al., 2006 (159)	Low	Unclear	Low	Unclear	Unclear	Low	Low	Low	Low
Emerman et al., 1991 (149)	Unclear	Unclear	Low	Low	Low	Low	Low	Low	Low
Crowe et al., 2010 (147)	High	Unclear	Low	Unclear	Unclear	Low	Unclear	Low	Low
Vaillancourt et al., 2009 (166)	High	Unclear	Low	Low	High	Low	Low	Low	Low
El-Solh et al., 1999 (148)	Low	Unclear	Low	Low	Low	Low	Low	Low	Unclear
Bojang et al., 2000 (142)	Unclear	Low	Low	Low	Low	Low	Low	Low	Low
Leibovici et al., 1991 (156)	Low	Unclear	Unclear	Low	Low	Low	Low	Low	Low
Stein et al., 2005 (165)	Unclear	Unclear	Low	Low	Unclear	Low	Low	Low	Low

*In studies where the CPR is applied retrospectively to data by the researcher using predictor data collected by the clinician, if there was no statement that researchers were blind to the reference standard the risk of bias was considered to be unclear. If predictor data was collected by the researcher and there was no statement that researchers were blind to the reference standard, the risk of bias was considered to be high; †When the reference standard comprised subjective tests, if there was no statement that those interpreting the reference standard test were blind to the results of either the CPR or clinician, the risk of bias was considered to be unclear; ‡If the method of determining disease status involved a combination of different test in which some test were applied to some patients and one test applied to all patients (differential verification) then the risk of bias was considered to be unclear. If performance of any of the reference standard test was dependent upon the results of the index test, the risk of bias was considered to be high. If it was not possible to determine whether all eligible patients had been included in the analysis the risk of bias was considered to be unclear. If it was clear that not all patients had been included in the analysis (due to missing outcome data or because data from the clinicians estimate or data necessary to derive the results of the CPR were not available (and these studies reported results for the comparison in different numbers of cases or only presented the results for cases on which data for both the comparisons was available, the risk of bias was considered to be high. Risk of bias was recorded as high if either of the issues relating to the reference standard test or analysis were high.

2.5.4 Study results

Results of the included studies are tabulated in Table 2.3 and Table 2.4, and presented graphically in Figure 2.3 and Figure 2.4.

There were 41 comparisons between CPRs and clinical judgment. (94, 138-144, 146-149, 152-165, 167). In 2 (5%) comparisons, (155, 156) CPRs reduced the proportion of missed diagnoses in those classified as not having the disease, but this was offset by classifying a larger proportion of study participants as having disease (more false positives). In 33 (80%) comparisons, (139, 142-144, 146, 148, 149, 152-154, 157, 161, 162, 165) the proportion of diagnoses missed by the CPR and clinical judgment was similar and in 7 of these comparisons (140, 142, 149, 158, 163, 164) CPRs classified a larger proportion of participants as not having disease (fewer false positives) and a similar proportion in 16. (139, 142-144, 146, 148, 149, 152-154, 157, 161, 162, 165) In 6 (15%) comparisons (94, 139, 141, 162, 167) the proportion of diagnoses missed by the CPR was greater. This was offset by classifying a smaller proportion of participants as having the disease (fewer false positives) in 2 (139, 162) comparisons. In 3 of the 6 comparisons (94, 141, 167) the CPRs classified a similar proportion of participants as having the disease. There was 1 comparison (94) where the CPR both missed more diagnoses and classified a larger proportion of participants as having the disease (more false positives), but no comparisons where the CPR missed fewer diagnoses and classified a larger proportion of participants as not having disease.

There were 5 comparisons between CPRs and the combination of CPR and clinical judgment. (145, 150, 151, 166) (Table 2.3, Table 2.4, Figure 2.3 and Figure 2.4) In 3 (60%) comparisons the proportion of diagnoses missed was similar (145, 151, 166) and in 2 (151, 166) of these comparisons, CPRs classified a larger proportion of study participants as not having disease (fewer false positives) than the combination of CPR and clinical judgment. In 2 (40%) comparisons, (145, 150) the proportion of diagnoses missed by the CPRs was greater while the proportion classified as not having disease by the CPRs and the combination of CPR and clinical judgment was similar. There were no comparisons between the combination of CPR and clinical judgment and clinical judgment alone.

There were 5 studies (139, 142, 143, 151, 154) of 10 comparisons, that used different thresholds for the CPR or clinical judgment (for example, Kabrhel et al., 2005 (154) compared clinical judgment to the Wells PE score at threshold <2 and ≤ 4). We report on the results of 9 of these comparisons, excluding the results of 1 comparison (151) where the proportions of interest (that is, the proportion classified as having disease or the proportion of missed

diagnoses) were similar at the different thresholds. This means that for a small number of comparisons (n=4) clinical judgment is counted twice. (139, 142, 143, 154)

Pulmonary embolism

From 9 studies in pulmonary embolism, there were 9 comparisons between the Wells PE score (original 3 level or 2 level score) and clinical judgment. (94, 143, 145, 153, 154, 162, 163) In 8 (89%) comparisons, (143, 145, 153, 154, 162, 163) the proportion of diagnoses missed by the score and clinical judgment was similar. In 1 of these, (163) the score classified a larger proportion of all participants as not having the disease (fewer false positives), a similar proportion in 5 comparisons (143, 145, 153, 154, 162) and a larger proportion of participants as having the disease (more false positives) in 2. (143, 154) In 1 (11%) comparison, (94) the proportion of diagnoses missed by the Wells PE score was greater, while the proportion of participants classified as not having the disease was similar. In 2 comparisons between the PERC Rule and clinical judgment, (155, 160) the rule reduced the proportion of missed diagnosis in 1, (155) but this was offset by classifying a larger proportion of participants as having the disease (more false positives). In the other comparison, (160) the proportion of diagnoses missed by the PERC rule and clinical judgment was similar. In 1 comparison (94) the Revised Geneva Score both missed more diagnoses and classified a larger proportion of participants as having the disease than clinical judgment. In 1 comparison (145) between the Geneva score and the combination of clinical judgment and score, the proportion of diagnoses missed by the CPR was greater.

Deep vein thrombosis

From 6 studies of DVT, there were 6 comparisons between the Wells DVT score and clinical judgment. (140, 141, 146, 158, 167) There were no comparisons in which the score reduced the proportion of missed diagnoses. In 4 (67%) comparisons the proportion of diagnoses missed by the score and clinical judgment was similar. (140, 146, 158) In 3 of these (141, 158) the score classified a larger proportion of all participants as not having disease (fewer false positives) and in 1 (146) the proportion was similar. In 2 comparisons (141, 167) the proportion of diagnoses missed by the CPR was greater, with a similar proportion classified as not having the disease. In 1 comparison (151) between the Oudega Rule and the combination of clinical judgment and Oudega Rule, the proportion of diagnoses missed was similar, with the rule classifying a larger proportion of participants as not having the disease (fewer false positives).

Streptococcal throat infection

There were 3 studies of streptococcal throat infection.

In 2 comparisons (144, 161) between the Centor Score (Modified and Original score combined with Tomkins Management Rule) and 1 comparison between the Walsh score and clinical judgment (144) the proportion of diagnoses missed and the proportion of all participants classified as not having disease was similar. In these studies clinicians would likely have been aware that all study participants would have pharyngeal swabs taken for testing as per study protocol. This may lead to an overestimate of the proportion of participants classified as not having disease by clinical judgment.

Foot and or ankle fracture

From 3 studies of foot and or ankle fracture, there were 3 (100%) comparisons between the Ottawa ankle and foot rules (OAR) and clinical judgment. (138, 152, 164) In all 3 comparisons the proportion of diagnoses missed by the CPR and clinical judgment was similar. In 1 of these (164) the rule classified a larger proportion of study participants as not having disease (fewer false positives) and in 2 comparisons (138, 152) the CPR classified a larger proportion of participants as having disease (more false positives). In the 2 comparisons from 2 studies (138, 152) in which the OAR classified a larger proportion of participants as having disease than clinical judgment, the clinicians when making a decision or diagnosis, would likely have been aware that all participants would be x-rayed as per study protocol (152) or would have known that an x-ray could be ordered at their discretion. (138) This may lead to an overestimate of the proportion of study participants classified as not having disease by clinical judgment.

Acute appendicitis

There were 2 studies of acute appendicitis. In 1 comparison (150) between the Fenyo Score and the combination of score and clinical judgment, the proportion of diagnoses missed by the score was greater while the proportion classified as not having disease was similar. In 1 comparison (157) between the Modified Alvarado Score and clinical judgment, the proportion of diagnoses missed and the proportion of all study participants classified as not having disease was similar.

Acute coronary syndrome, pneumonia, head injury in children, cervical spine injury, active pulmonary tuberculosis, malaria, bacteraemia and influenza.

Of 8 studies (11 comparisons) addressing a variety of conditions, the CPRs showed either an improvement in the proportion of missed diagnosis or the proportion classified as not having disease, but this was often offset by a worsening of the other measure.

Table 2.3. Characteristics and results of included studies for conditions with >2 studies

Study	Setting	Method of establishing status of target disorder	prevalence (n/N)	Comparison (method of estimating the probability of target disorder, making a diagnosis or management decision)	Threshold (low risk if)	Sensitivity (95% CI)	Specificity (95% CI)	% missed cases of disease among those classified as not having disease	% classified as not having disease
Pulmonary embolism									
Runyon et al., 2005 (162)	ED	Medical record review, F/U by mail or telephone and death records at 1.5 months	6% (144/2477)	Clinical judgment + structured data collection Wells PE score calculated by researcher Charlotte score calculated by clinician Clinical judgment alone	<15% <2 Safe <15%	69 (61-76) 62 (54-70) 36 (28-45) 69 (65-73)	72 (70-74) 75 (73-77) 89 (88-91) 70 (69-71)	2.6 (1.9-3.5) 3.0 (2.3-3.9) 4.2 (3.5-5.2) 3.1 (2.7-3.6)	69 (67-71) 73 (70-74) 88 (87-89) 68 (67-69)
Kabriel et al., 2009 (153)	ED	Adjudicated review of imaging results, medical records and F/U at 1.5 months	7% (545/7940)	Wells PE score calculated by researcher	<2	68 (64-72)	72 (71-73)	3.2 (2.7-3.7)	69 (68-70)
Kline et al., 2008 (155)	ED	Adjudicated review of imaging results, medical records and F/U at 1.5 months	7% (561/8138)	Clinical judgment + structured data collection PERC Rule calculated by researcher	<15% No criteria present	71 (67-74) 96 (94-97)	69 (68-71) 25 (24-26)	3.0 (2.6-3.5) 1.3 (0.9-1.9)	67 (66-68) 24 (23-25)
Kabriel et al., 2005 (154)	ED	Review of medical records at 3 months F/U	10% (61/607)	Clinical judgment + structured data collection Wells PE score calculated by researcher	Alternate diagnosis not less likely <2	54 (41-67) 79 (66-88)	76 (73-80) 57 (53-61)	6.3 (4.4-8.9) 4.0 (2.4-6.7)	73 (70-77) 54 (50-58)
Carrier et al., 2006 (143)	NMD	Patient follow-up by telephone or return appointment at 3 months	18% (76/413)	Clinical judgment + structured data collection Wells PE score calculated by researcher	≤4 <20%	59 (46-72) 86 (76-93)	78 (74-81) 38 (33-43)	5.5 (3.8-8.1) 7.5 (4.3-13.0)	74 (70-77) 34 (30-38)
Chagnon et al., 2002 (145)	ED	Follow-up (method not specified) at 3 months	26% (71/277)	Rodgers model calculated by researcher Clinical judgment + access to the Geneva Score Wells PE score calculated by researcher	No criteria present 'low' <2	95 (87-99) 83 (73-91) 73 (61-83)	19 (15-24) 41 (35-46) 69 (63-76)	5.8 (2.3-14.0) 8.7 (5.1-14.3) 11.7 (7.6-17.6)	17 (13-21) 36 (32-41) 59 (53-64)
Penaloza et al., 2012 (160)	ED	Review of imaging results, medical records and patient or relative follow-up at 3 months	30% (286/959)	Geneva Score calculated by clinician Clinical judgment + structured data collection Revised Geneva Score + PERC Rule calculated by researcher	≤4 'low' <4 and no criteria present	72 (60-82) 91 (87-94) 99 (97-99.6)	64 (57-71) 55 (52-59) 9 (7-12)	13.2 (8.7-19.5) 6.5 (4.5-9.4) 6.2 (2.4-14.8)	55 (49-61) 42 (39-45) 7 (5-9)
Sanson et al., 2000 (163)	IP, OPD	Perfusion lung scintigraphy or pulmonary angiography	31% (160/517)	Clinical judgment + structured data collection Wells PE score calculated by researcher	No criteria present <20%	99 (97-99.6) 91 (85-96)	10 (8-13) 16 (12-21)	5.4 (2.1-13.1) 19.0 (10.9-30.9)	8 (6-10) 14 (11-18)
Penaloza et al., 2013 (94)	ED	Review of imaging results, medical records and patient or relative follow-up at 3 months	31% (325/1038)	Clinical judgment + structured data collection Wells PE score calculated by researcher Revised Geneva Score calculated by researcher	'low' <2 <4	90 (86-93) 82 (77-85) 89 (85-92)	58 (54-61) 60 (56-63) 33 (30-37)	7.6 (5.5-10.5) 12.6 (9.9-15.8) 13.0 (9.5-17.5)	43 (40-46) 47 (44-50) 26 (23-29)

Study	Setting	Method of establishing status of target disorder	prevalence (n/N)	Comparison (method of estimating the probability of target disorder, making diagnosis or management decision)	Threshold (low risk if)	Sensitivity (95% CI)	Specificity (95% CI)	% missed cases of disease among those classified as not having disease (95% CI)	% classified as not having disease (95% CI)
Deep vein thrombosis									
Geersing et al., 2010 (151)	PC	Clinical probability, ultrasound and F/U at 3 months	14% 136/1002	Clinical judgment + access to the Oudega Rule	<10%	98 (94-99)	24 (22-27)	1.4 (0.5-4.0)	21 (19-24)
Bigaroni et al., 2000 (140)	ED, OPD	D-dimer, ultrasound, other imaging and telephone F/U at 3 months	17% (28/165)	Oudega Rule calculated by clinician Clinical judgment + structured data collection Wells DVT score calculated by junior clinician Wells DVT score calculated by senior clinician	<=3 'Low risk' <1 <1	95 (90-98) 98 (85-99.8) 71 (53-85) 79 (61-90)	57 (54-60) 46 (38-54) 75 (67-82) 74 (66-80)	1.4 (0.6-2.9) 0.0 (0.0-5.8) 7.2 (3.7-13.6) 5.6 (2.6-11.7)	50 (47-53) 38 (31-46) 67 (60-74) 65 (57-72)
Miron et al., 2000 (158)	ED, OPD	D-dimer, ultrasound, other imaging and telephone F/U at 3 months	21% (57/270)	Clinical judgment + structured data collection Wells DVT score calculated by researcher	<20%	98 (91-99.7)	36 (30-43)	1.3 (0.2-6.9)	29 (24-35)
Blattler et al., 2004 (141)	OPD	Ultrasound and telephone F/U at 6+ months	28% (57/206)	Clinical judgment + structured data collection Wells DVT score calculated by researcher	<1 'Low risk'	93 (83-97) 81 (69-89)	57 (50-63) 85 (79-90)	3.2 (1.3-7.9) 8.0 (4.5-13.7)	46 (40-52) 67 (60-73)
Cornuz et al., 2002 (146)	VL	Ultrasound, other imaging, mail or telephone F/U	29% (82/278)	Clinical judgment + structured data collection Wells DVT score calculated by researcher	Low <20%	54 (42-67) 87 (78-92)	84 (77-89) 38 (32-45)	17.2 (12.0-24.0) 12.8 (7.3-21.5)	73 (67-79) 31 (26-37)
Wang et al., 2013 (167)	OPD	Ultrasound and telephone or email F/U at 1.5 months	47% (191/405)	Wells DVT score calculated by researcher Clinical judgment alone Wells DVT score** calculated by clinician	<1 'Safe' <=1	83 (73-90) 76 (70-82) 62 (55-69)	49 (42-56) 89 (84-92) 72 (66-78)	12.8 (7.8-20.4) 19.2 (14.6-24.7) 31.9 (26.1-38.2)	39 (34-45) 58 (53-63) 56 (51-61)
Streptococcal throat infection									
Cebul and Poses, 1986 (144)	PC/ Adults	Throat culture	5% (15/310)	Clinical judgment + structured data collection Walsh model + Tomkins management rule calculated by researcher Centor model + Tomkins management rule calculated by researcher	No treatment No treatment No treatment	53 (27-79) 80 (52-95) 100 (75-100)	68 (62-73) 57 (51-72) 66 (60-72)	3.4 (1.7-6.9) 1.5 (0.5-4.3) 0.0 (0.0-2.2)	67 (61-72) 65 (59-70) 63 (57-69)
Rosenberger et al., 2002 (161)	ED/ Mixed	Pharyngeal swab culture	25% (32/126)	rule calculated by researcher Clinical judgment + structured data collection + rapid test Modified Centor Score calculated by researcher	No treatment No treatment No treatment	90 (74-98) 97 (84-99.5)	92 (87-95) 78 (68-86)	5.0 (2.0-12.2) 1.4 (0.2-7.3)	64 (55-71) 59 (50-67)

Table 2.3. Continued

Study	Setting	Method of establishing status of target disorder	prevalence (n/N)	Comparison (method of estimating the probability of target disorder, making diagnosis or management decision)	Threshold (low risk if)	Sensitivity (95% CI)	Specificity (95% CI)	% missed cases of disease among those classified as not having disease (95% CI)	% classified as not having disease (95% CI)
Streptococcal throat infection continued									
Attia et al., 2001 (131)	ED, OPD/ Children	Tonsillopharyngeal swab culture	37% (218/587)	Clinical judgment + structured data collection Clinical prediction rule of Attia calculated by researcher	<=50% 0 ≤3	72 (66-78) 99 (97-99.9) 18 (13-24)	60 (55-65) 5 (3-7) 97 (95-99)	21.6 (17.2-26.7) 11.8 (3.3-34.3) 34.7 (30.7-39.0)	48 (44-52) 3 (2-5) 91 (89-94)
Ankle and or foot fracture									
Glas et al., 2002 (144)	ED/ Adults	Ankle and mid-foot x-ray	6% (41/647)†	Clinical judgment + structured data collection OAR – ankle and foot calculated by researcher Leiden ankle rule calculated by researcher	No X-ray Negative ≤7	98 (87-99.6) 98 (87-99.6) 88 (74-96)	66 (62-70) 26 (22-29) 57 (53-61)	0.3 (0.0-1.4) 0.6 (0.1-3.5) 1.4 (0.6-3.3)	62 (59-66) 24 (21-28) 55 (51-58)
Singh-Ranger and Marathias 1999 (156)	ED/ Adults	Ankle x-ray	17% (3/18)§	Clinical judgment alone OAR – ankle calculated by researcher	No fracture Negative	100 (31-100) 100 (31-100)	0.0 (0.0-22) 67 (38-85)	0.0 (0.0-0.0) 0.0 (0.0-27.8)	0 (0-18) 56 (34-75)
Al Omar and Baldwin 2002 (130)	ED/ Children	Ankle or mid-foot x-ray	21% (17/80) †‡	Clinical judgment + structured data collection OAR – ankle and foot calculated by researcher	No fracture Negative	65 (38-86) 100 (80-100)	76 (64-86) 30 (19-43)	11.1 (5.2-22.2) 0.0 (0.0-16.8)	68 (57-77) 24 (16-34)

ED – emergency department; OPD – Outpatient department; VL – vascular laboratory; PC – primary care; NMD – nuclear medicine department; F/U – follow-up; OAR – Ottawa ankle rules; PE – pulmonary

embolism; DVT – deep vein thrombosis

*% missed cases of disease among those classified as not having disease (FN/FN+TN or 1-negative predictive value); †% classified as not having disease (FN+TN/total study N); ‡ankle or mid-foot fracture; § ankle

fracture; || Includes Salter Harris fractures; **2-category Wells DVT score

Table 2.4. Characteristics and results of included studies for conditions with ≤ 2 studies

Study	Setting	Method of establishing disease status	Prevalence (n/N)	Comparison (method of estimating probability, making diagnosis or management decision)	Threshold (low risk if)	Sensitivity (95%CI)	Specificity (95% CI)	% missed cases of disease among those classified as not having disease (95% CI)*	% classified as not having disease (95% CI)†
Fenyo, 1987 (150)	IP	Intraop diagnosis, histopathology of excised appendices and record review at 1-2 years	31% (256/830)	Clinical judgment + access to results of Fenyo Score Fenyo Score calculated by researcher	No surgery ≤ 11	100 (99-100) 90 (86-94)	91 (88-93) 92 (89-94)	1.0 (0.0-0.7) 4.6 (3.1-6.6)	63 (59-66) 66 (63-69)
Meltzer et al., 2013 (157)	ED	Surgical pathology, CT scan or telephone F/U at 7 days	20% (53/261)	Clinical judgment alone Modified Alvarado score calculated by researcher	Appendicitis not most likely diagnosis < 4	79 (66-89) 72 (58-83)	68 (61-74) 54 (47-61)	7.2 (4.1-12.5) 11.8 (7.3-18.6)	58 (52-64) 49 (43-55)
Mitchell et al., 2006 (159)	ED	Review of medical records and telephone F/U at 1.5 months	5% (51/1114)	Clinical judgment + structured data collection ACI-TIPI calculated by researcher	$\leq 2\%$ $\leq 2\%$	96 (87-99) 100 (93-100)	27 (25-30) 5 (4-7)	0.7 (0.2-2.5) 0.0 (0.0-6.4)	26 (24-29) 5 (4-7)
Emerman et al., 1991 (149)	ED, OPC	Posteroanterior and lateral chest x-ray	7% (21/290)	Clinical judgment + structured data collection Diehr Score calculated by researcher Heckerling Score calculated by researcher Gennis Rule calculated by researcher	No radiograph ≤ 0 < 2 No variable present	86 (64-97) 67 (43-85) 71 (48-89) 62 (39-82)	58 (52-64) 67 (61-73) 67 (61-73) 76 (70-81)	1.9 (0.7-5.4) 3.7 (1.8-7.5) 3.3 (1.5-6.8) 3.8 (1.9-7.2)	55 (49-60) 65 (59-70) 65 (59-70) 74 (68-78)
Crowe et al., 2010 (147)	ED	Medical record review of imaging tests, observation and readmission	7% (73/1065)	Singal Score calculated by researcher Clinical judgment alone CHALICE criteria calculated by researcher	Probability < 0.26 No CT scan No criteria present	76 (53-92) 95 (87-98) 89 (80-95)	55 (49-61) 86 (84-88) 57 (54-60)	3.3 (1.4-7.4) 0.5 (0.2-1.2) 1.4 (0.7-2.7)	53 (47-58) 81 (78-83) 54 (51-57)
Vaillancourt et al., 2009 (166)	ES	Radiographic imaging and telephone or mail F/U at 14 days	1% (12/1974)	Clinical judgment + access to results of Canadian C-Spine Rule Canadian C-Spine Rule score calculated by researcher	Negative Negative	100 (73-100) 100 (73-100)	38 (36-40) 43 (40-45)	0.0 (0.0-0.5) 0.0 (0.0-0.6)	38 (35-40) 43 (40-45)




Table 2.4. Continued

Study	Setting	Method of establishing disease status	Prevalence (n/N)	Comparison (method of estimating probability, making diagnosis or management decision)	Threshold (low risk if)	Sensitivity (95%CI)	Specificity (95% CI)	% missed cases of disease among those classified as not having disease (95% CI)*	% classified as not having disease (95% CI)†
El Solh et al., 1999 (148)	IP	Culture of respiratory specimens	9% (11/119)	Clinical judgment + structured data collection El Solh rule calculated by researcher	No active TB Negative	64 (31-89) 100 (71-100)	79 (70-86) 69 (60-78)	4.5 (1.8-11.0) 0.0 (0.0-4.9)	75 (66-82) 63 (54-71)
Bojang et al., 2000 (142)	OPD	Temperature and parasitemia on blood film	35% (133/382)	Clinical judgment alone Olaeye algorithm calculated by researcher	No malaria <7	82 (74-88) 90 (83-94)	61 (55-67) 63 (57-69)	13.6 (9.3-19.5) 8.2 (4.9-13.3)	46 (41-51) 46 (41-51)
Leibovici et al., 1991 (156)	IP	Blood culture	14% (36/257)	Clinical judgment + structured data collection Rule of Leibovici calculated by researcher	<8 No bacteremia	90 (83-95) 53 (36-70)	78 (72-83) 84 (79-89)	17.1 (12.8-22.4) 8.5 (5.4-13.2)	61 (56-66) 79 (74-84)
Stein et al., 2005 (165)	ED	Reverse transcriptase PCR assay for influenza A and B	21% (53/258)	Clinical judgment alone Cough and fever rule calculated by researcher	<20% No influenza Negative	97 (85-99.5) 29 (17-44) 41 (27-57)	60 (53-67) 92 (87-95) 92 (87-95)	0.8 (0.1-4.2) 18.0 (13.2-24.1) 14.8 (10.4-20.7)	52 (46-58) 87 (82-91) 84 (78-88)

*% missed cases of disease (FN/FN+TN or 1-NPV); †% classified as low risk (FN+TN/total N)

Figure 2.3. Results of the included studies for conditions with >2 studies

Bar chart interpretation

Clinical judgment  Clinical judgment with access to a CPR  CPR 
A = Negative 'test' B = Negative 'test' and no disease C=Positive 'test' D=Positive 'test' and disease

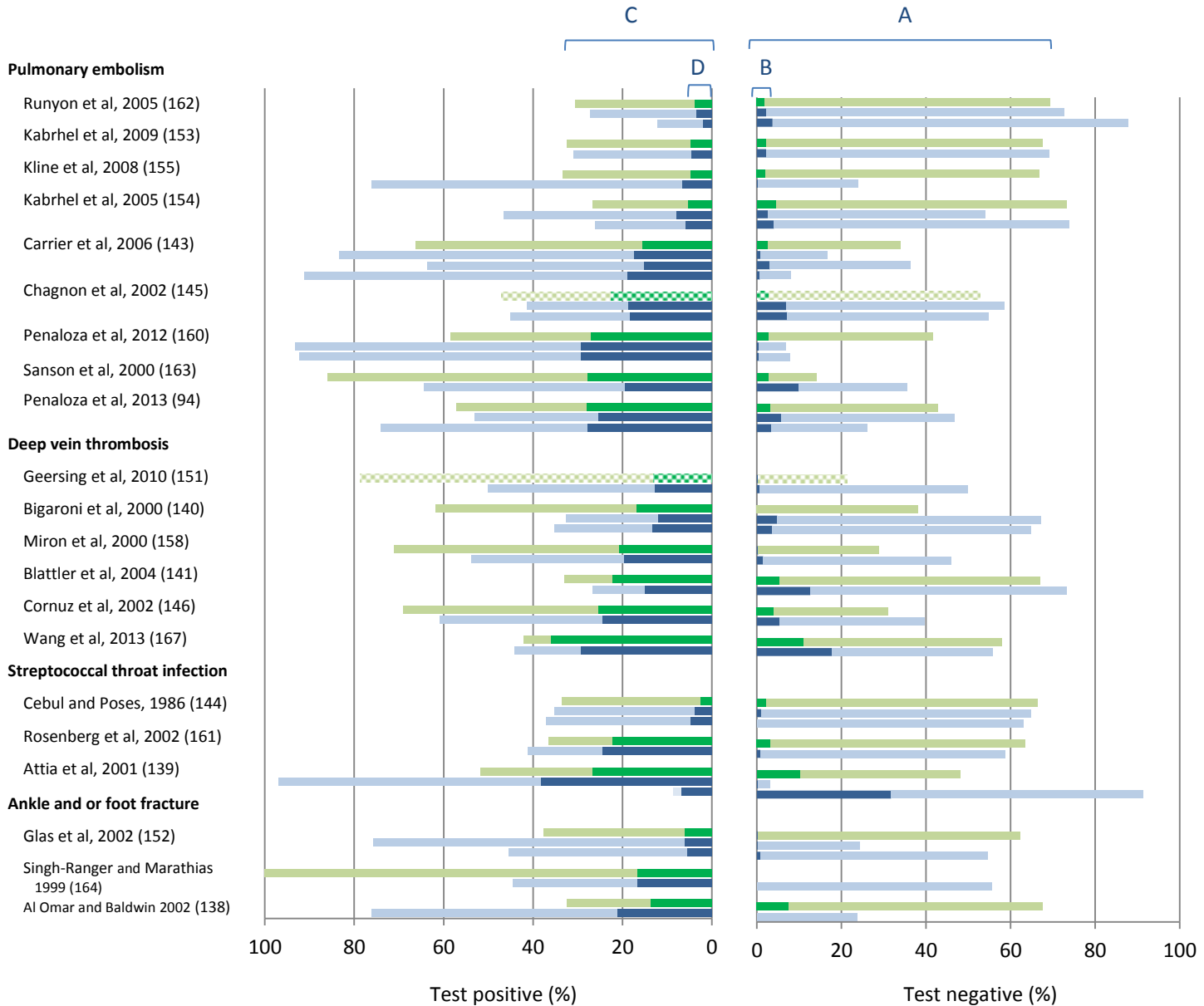


Figure 2.4. Results of the included studies for conditions with ≤ 2 studies

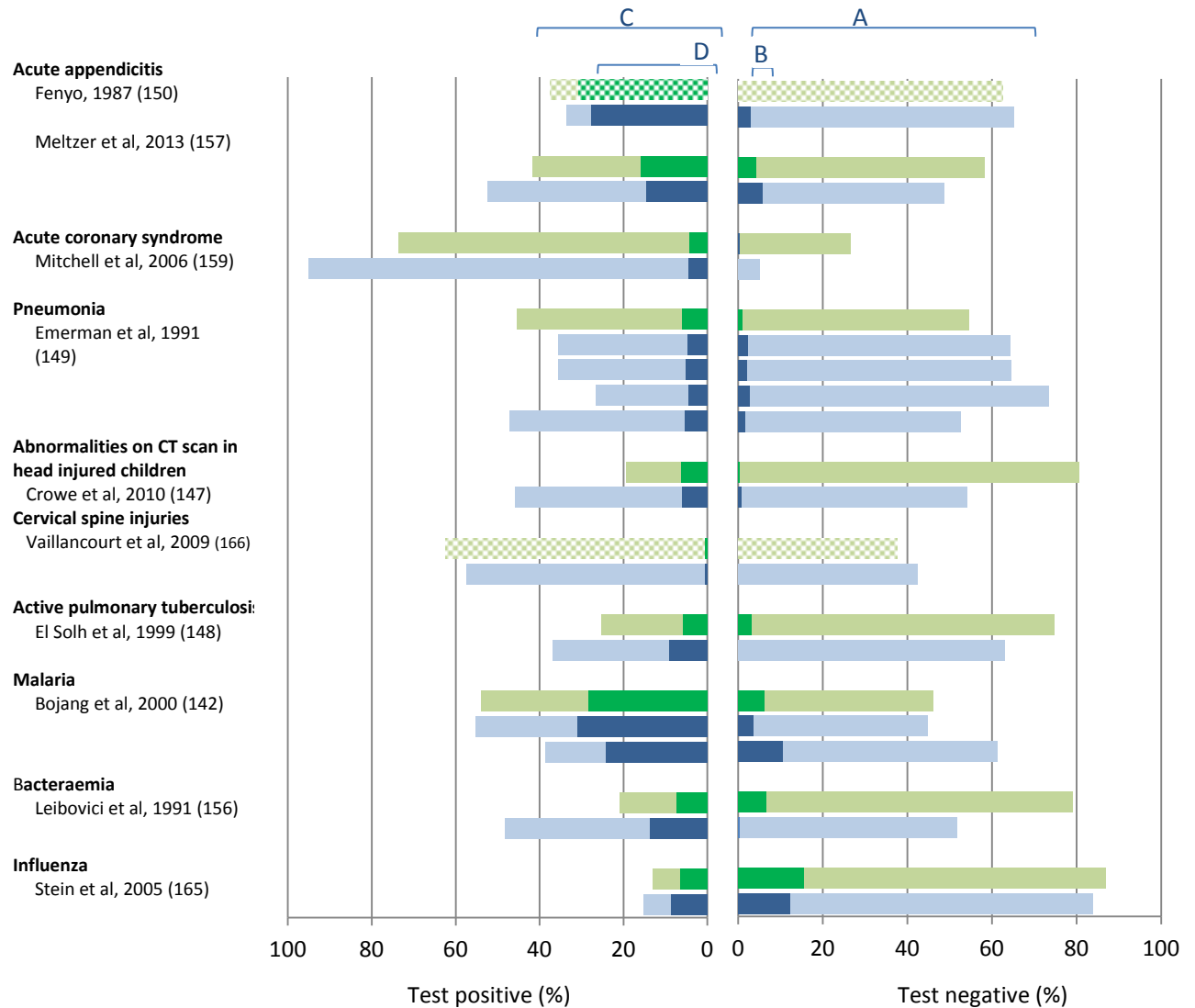
Bar chart interpretation

Clinical judgment

Clinical judgment with access to a CPR

CPR

A = Negative 'test' B = Negative 'test' and no disease C = Positive 'test' D = Positive 'test' and disease



2.6 Discussion

In this review, CPRs were rarely superior to clinical judgment and there was generally a trade-off between the proportion of study participants classified as not having disease and among those classified as not having disease, the proportion of missed diagnoses of disease. CPRs for the diagnosis of DVT generally classified a larger proportion of all participants as not having disease than clinical judgment, but this was often at the expense of missed diagnoses. In other disease areas, CPRs showed either an improvement in the proportion classified as not having disease or the proportion of missed diagnoses, but often with the trade-off of worsening the other measure. These findings, however, are limited by the small number of studies for many of the conditions, the design features and generally unclear or high risk of bias in many of the included studies.

Trade-offs in the proportion classified as not having disease and the proportion of missed diagnosis by CPRs and clinical judgment seen in this review probably represent differences in the diagnostic threshold for positivity of the two judgment methods. For example, CPRs might be developed to avoid missing people with disease and as such the threshold for positivity is set very low. The CPR would therefore likely be safer than clinical judgment where the threshold for positivity is implicitly set and variable between and within clinicians, but this is often at the expense of classifying fewer participants as not having disease (and thereby avoiding further testing or treatment). Whether clinical judgment or a CPR is the preferred judgment methods for a particular clinical condition will therefore depend on the relative benefits and harms arising from true positive and false positive diagnosis.

Variability in the proportion classified as not having disease and proportion of missed diagnoses of CPRs compared with clinical judgment, even amongst studies of the same CPR, may be explained in part by features of the clinical setting of the studies. Differences in study design and methodology, including the type of CPR tested (logistic regression model or other statistical technique), the rigour with which it was developed, the case-mix of the study population, 'modifications' to clinical judgment (with or without structured data collection), by whom (novice or experienced clinicians) or the way in which the result of the CPR is derived (calculation by clinician or researcher) may also explain the variation in performance seen in the studies included in this review. In many studies, clinicians collected diagnostic data on a structured data collection form. This systematic collection of diagnostic information may improve the observed diagnostic accuracy of the clinicians. (168) Clinician experience has also been shown to improve the accuracy of diagnosis. (169)

Variability in the outcomes of clinical judgment and CPRs within conditions may also be explained by the method used to elicit clinical judgment, as the method used will likely be associated with the implicit threshold for positivity. In studies of appendicitis for example, clinical judgment was a clinician's diagnosis of appendicitis or the clinician's actual action to perform surgery or not. In studies of ankle fracture, clinical judgment was either a clinician's diagnosis of fracture or their intention to x-ray a patient, and for studies of sore throat, clinical judgment may have been a clinician's actual action to prescribe antibiotics or not, or a clinician's statement of their intention to treat with antibiotics. The clinician's threshold for positivity will likely be higher for instance, if asked to provide a diagnosis (diagnostic threshold) than when asked of their intention to do further definitive testing (testing threshold). Where clinical judgment was elicited by obtaining a clinician's probability estimate on a continuous scale, there was also variation in the thresholds applied by study researchers. For studies of pulmonary embolism for example, thresholds were applied at probabilities of 15 or 20%.

The design of the studies included in this review allows comparison of the performance of CPRs and clinical judgment when applied independently. In practice, however CPRs are likely to be used as tools to support or complement clinical judgment. When used in this manner, the performance of the diagnostic CPRs may vary from that shown in this review. The effect of a CPR when used in conjunction with clinical judgment can only be fully tested in a study design in which participants are assigned (ideally randomly) to apply or receive clinical judgment alone or clinical judgment with access to a CPR. However, studies of diagnostic accuracy or incremental value (12, 32) provide a useful and less costly interim step in the evaluation of CPRs prior to a randomised controlled trial and can guide future research.

Our study shows that, in the context of medical diagnosis, CPRs do not consistently classify more individuals as not having disease or miss fewer diagnoses among those classified as not having disease than clinical judgment. This is in contrast to several reviews comparing clinical and statistical methods of prediction, often combining studies from fields as diverse as education, criminology and healthcare, which have generally found statistical methods to be superior. (36, 55, 58) A more recent body of research however has found that when formally tested, heuristics, proposed as models of human judgment are, in some situations as accurate as, or more accurate than statistical models. (56) A review comparing the diagnostic accuracy of doctors and statistical tools for acute appendicitis (170) found that statistical tools had greater specificity than clinicians. However, most of the studies included in this review were excluded from the present review because a) the statistical tools and clinical judgment were

not applied at the same time point or b) the statistical tools and clinical judgment were not applied to the same participants.

Due to variation in the design and purpose of the included studies, we did not attempt meta-analysis across or within study conditions. Instead, we compared CPRs and clinical judgment using two measures 1) the proportion of all study participants classified as not having disease (a measure of efficiency) and 2) the proportion of participants among those classified as not having disease, who actually have the disease (false negative rate, a measure of safety).

Because many CPRs seek to either improve diagnosis or identify a group of patients who do not require additional testing, we believe these are the most clinically relevant measures.

Though these measures are dependent on the prevalence of the disease in the study population, the studies were judged to have been undertaken in relevant clinical settings.

Traditional measures of diagnostic accuracy, such as sensitivity, specificity and area under a receiver operator characteristic curve are often favoured accuracy metrics because they are commonly believed to be unaffected by disease prevalence, though this has recently been shown not to be the case. (78) The proportion of participants classified as having disease and the proportion with false positive results can also be obtained from Figure 2.3 and Figure 2.4 and the traditional measures of diagnostic accuracy from Table 2.3 and Table 2.4.

The majority of included studies were judged to be at high or unclear risk of bias on 2 or more of the 4 risk of bias domains assessed. Differential verification (the results of clinical judgment or the CPR influence the performance of reference tests) and incorporation bias (the results of the CPR are used to make the final diagnosis) affected many studies, particularly studies of DVT and PE. Further, studies commonly did not include all eligible cases in the analysis and often it was not clear whether researchers applying a CPR retrospectively to a dataset were blind to the results of the reference standard. The design of studies of ankle fracture and streptococcal throat infection may also have led to inaccurate estimates of the diagnostic accuracy of clinical judgment. In these studies, the clinicians' diagnosis or decision that x-ray or antibiotics are necessary may have been influenced by knowledge that all or most study participants would undergo confirmatory testing with an x-ray or throat swab. In this review, in two of the three studies of ankle and or foot fracture, the Ottawa Ankle Rules were considerably less efficient than clinical judgment that a fracture was present or that an x-ray was necessary. This finding conflicts with that a multicentre randomised controlled trial in which application of the rules lead to x-rays for 79% of study participants compared to 99.6% of participants when the decision was made by emergency department physicians. (171)

The database searches to identify studies for the review were conducted up to March 2013 and eligible studies may have been published since this time. Because of the size of the search, not all titles and abstracts identified in electronic searches were screened by 2 reviewers. However, a second reviewer screened a subset of titles and did not find any additional studies. The search terms used may not have located all eligible studies, but manual searches of systematic reviews of CPRs and comprehensive reference and citation checking minimise this possibility. An assessment of the risk of bias in the studies deriving the CPRs or the 'useability' features of the CPRs evaluated in this review was not conducted, but updates to this review should seek to do this. Such information may assist in the interpretation of the results of the review.

While CPRs show promise as a way of improving clinical decision making, to date there have been limited studies comparing, in the same participants, the accuracy of CPRs and clinical judgment, and those studies often had design issues that raised the potential for bias and made interpretation of their results difficult. Though detailed guidance on the validation and evaluation of prediction models and rules is available, (42, 105) guidance on issues specific to studies comparing the diagnostic performance of CPRs and clinical judgment may improve this situation. To inform of the potential of diagnostic CPRs to improve diagnosis and patient outcomes when the CPR is used in combination with clinical judgment, particularly in situations where the clinician has a high degree of uncertainty, an analysis of studies comparing care provided when clinicians have access to a diagnostic CPR with usual care would be useful.

2.6.1 Conclusion

The limited studies included in this review show that none of the CPRs evaluated to date are clearly superior to clinical judgment across a range of medical conditions. They also show variation in the comparative performance of clinical judgment and CPRs between studies for the same condition and between the same CPRs. There is generally a trade off in the proportion classified as not having disease and missed diagnosis that is most likely due to different thresholds for positivity associated with clinical judgment and CPRs. The current review highlights some of the methodological issues relating to the conduct of studies comparing CPRs and clinical judgment, with design features of many of the included studies increasing the potential for bias.

2.6.2 Acknowledgements

We would like to thank Rae Thomas for help with screening of the titles and abstracts.

2.6.3 Author contributions

Conceived the experiment: SS. Designed the experiment: SS, JD and PG. Analysed the data: SS and JD. Wrote the first draft of the manuscript: SS. Contributed to the writing of the manuscript: SS, JD, PG. Agree with manuscript results and conclusions: SS, JD, PG.

International Committee of Medical Journal Editors (ICMJE) criteria for authorship read and met: SS, JD, PG.

SS received funding from an Australian Postgraduate Award scholarship and the Screening and diagnostic Test Evaluation Program which is supported by a National Health and Medical Council Program Grant (<https://www.nhmrc.gov.au/>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.6.4 Modifications to the published journal article that appear in this chapter

Location	Alteration/Addition
2.4.4 Data synthesis and analysis, page 42 of the thesis	Addition: When the output of the CPR or clinical judgment was not a binary decision or action (e.g. the CPR classified individuals as low, moderate or high risk, or clinical judgment was a clinicians' decision to not test or treat, to test, or to test and treat), we dichotomised the output by combining probability estimates that were not 'low' (e.g. moderate and high or possible or probable), and decision or actions that involved tests or treatment (e.g. for a score throat score, the directive output 'culture' and 'culture and treat' were combined and compared to 'no culture or treatment').
2.4.4 Data synthesis and analysis, page 43 of the thesis	Addition: Box 2.1
2.4.4 Data synthesis and analysis, page 43 of the thesis	Alteration: If studies reported different thresholds for clinical judgment or the CPR, and if the proportions (i.e. those classified as not having disease and the proportion of missed diagnoses) at the different thresholds were similar both in favour of, or opposed to the CPR or clinical judgment at the different thresholds (this only occurred in 1 study included in this review) we reported only the comparison for the threshold with the highest Youden's index ((sensitivity + specificity)-1).

Chapter 3

The effects of care provided with and without diagnostic prediction rules on patient and process outcomes

Sanders S, Rathbone J, Bell K, Glasziou P, Doust J. A systematic review of the effects of diagnostic clinical prediction rules. Submitted to BMJ Open May 2016.

3.1 Preface to Chapter 3

Decisions on whether to implement a diagnostic prediction tool in practice should be informed by rigorous evidence that they do patients more good than harm. This chapter presents the second study addressing the thesis aim to determine the value of diagnostic prediction rules as tools to assist clinical judgment. It does so by providing a systematic review of randomised trials comparing the effect of care provided with and without a diagnostic prediction rule on patient and process outcomes. The review also evaluates the reporting of the study interventions and implementation methods.

3.2 Abstract

Background: Diagnostic clinical prediction rules (CPRs) are developed to improve clinical decision making with the expectation that this will lead to either improved patient outcomes or other healthcare benefits such as reduced resource use, without adversely affecting patients. We conducted a systematic review to assess the effects of diagnostic CPRs on process of care and patient outcomes.

Methods and findings: We searched MEDLINE and CENTRAL with citation and reference checks in Web of Science, for randomised trials comparing a diagnostic strategy incorporating a diagnostic CPR, with a diagnostic strategy without a CPR.

Twenty five studies evaluating diagnostic CPRs for 14 different conditions were included with most evaluating a management decision as the primary study outcome. The majority of the studies were judged to be at high or uncertain risk of bias on three or more of the six domains of bias, one of which was performance bias.

Exposure to a diagnostic CPR for Group A Streptococcus throat infection reduced symptoms (1 study) and antibiotic prescriptions (5 studies, RR 0.86 95%CI 0.75 to 0.99). For cardiac chest pain, a diagnostic strategy incorporating a CPR improved early discharge rates (1 study) and decreased hospitalisations (1 study). The Ottawa Ankle Rules reduced radiography requests when used in conjunction with clinical examination (1 study) but had no effect on length of stay as a triage test prior to clinical examination (1 study). CPRs for acute appendicitis reduced time to therapeutic operation (1 study) and nontherapeutic operations but the effect was not statistically significant (5 studies, pooled RR 0.68 95%CI 0.43 to 1.08). CPRs in 3 studies of children with fever and possible serious bacterial infection did not improve process outcomes (prescribing, length of stay, appropriate test use). Details of study interventions and implementation were infrequently reported.

Conclusion: Diagnostic CPRs had a positive effect on process outcomes in some clinical conditions but few studies evaluated their effects on patient outcomes. These results may be context specific however and future studies should seek to provide detail on how the CPR might alter the diagnostic pathway, measure relevant patient and process outcomes, and to improve reporting of the study interventions and the way in which the CPR is implemented in practice.

3.3 Introduction

Diagnostic clinical prediction rules (CPRs) are tools intended to supplement clinicians' diagnostic reasoning and judgment (42) by providing an estimate of the probability of the presence of a particular disease in an individual, and/or by suggesting a course of clinical action based on the underlying probability estimate.

The decision to introduce a diagnostic CPR into practice should ideally be based on evidence that actual use leads to either 1) improved patient outcomes or 2) other benefits such as reduced resource use, relative to the current alternative pathway, without adversely affecting patients. The vast majority of studies of diagnostic CPRs in the literature however, have focused on establishing the accuracy of the CPR relative to a reference standard test in derivation and validation studies, with no comparison to the existing diagnostic pathway. Often this information is used to decide on the clinical usefulness of the CPR. However, diagnostic accuracy does not necessarily translate into patient benefits, (172) nor is it a necessary prerequisite for improved patient health as a CPR may alter patient health through other non-decisional routes including by changing the timing of decisions and actions relative to the existing pathway, or through direct effects of the CPR itself. (50) Therefore impact studies are necessary. These studies compare testing strategies with and without a diagnostic CPR reporting relevant patient and/or process outcomes. (5)

Whether implementation of current, validated diagnostic clinical prediction rules leads to more benefit than harm is unclear. The effects of CPRs as part of a broader group of clinical decision support tools (computerised and non-computerised tools for improving clinical decision making including, among other things, prediction rules, guideline based recommendations, alerts or reminders, condition specific order sets and contextually relevant reference information) have been extensively reviewed. (173-177) However, the effect of prediction rules specifically is difficult to discern, as these reviews have not analysed effects according to type of clinical decision support system implemented. To our knowledge, there has been no review of the effect of diagnostic prediction rules developed for a range of conditions commonly encountered in clinical medicine. Such a review may inform the selection and implementation of diagnostic CPRs in practice and future CPR research.

To determine the effect of exposure to diagnostic CPRs on patient and process outcomes, we reviewed studies randomly allocating clinicians or patients to care provided with a diagnostic CPR, or to care without a diagnostic CPR.

3.4 Methods

This review was performed following methods detailed in the systematic review protocol and is reported in line with the PRISMA statement for reporting in systematic reviews and meta-analyses (Appendix B).

3.4.1 Data sources and searches

We searched electronic databases including MEDLINE and The Cochrane Central Register of Controlled Trials (CENTRAL) to June 2015 using MeSH and text word terms for the intervention and a study design filter (Appendix B). We checked systematic reviews of diagnostic clinical prediction rules and clinical decision support systems identified using PubMed Clinical Queries. Reference lists of studies obtained in full text were checked and studies included in the review were forward searched using the Science Citation Index Expanded in Web of Science. The International Clinical Trials Registry Platform (ICTRP) was searched (June 2015) to identify trials planned, in progress or recently completed.

3.4.2 Study selection

We included randomised controlled studies allocating clusters of individuals, or individual clinicians or patients, to a group 'exposed' to a diagnostic strategy comprised of or incorporating a previously derived diagnostic clinical prediction rule (experimental), or to care provided without a CPR (control).

Eligible experimental interventions comprised the provision of a diagnostic CPR or the output of it, or a diagnostic strategy incorporating a diagnostic CPR (for example a strategy including a CPR and another laboratory or imaging test) to a clinician. A diagnostic clinical prediction rule was defined as a combination of variables obtained from history, examination or diagnostic testing, developed using a statistical method and which provide a probability of the presence of disease for an individual and/or suggested a diagnostic or therapeutic course of action. The course of action may relate to further testing or management or both. Studies evaluating tools incorporating a CPR designed for use by the patient or as part of joint decision making by the clinician and patient were not eligible for inclusion. The control intervention was an alternative diagnostic test or testing pathway that did not incorporate a diagnostic CPR. Studies that reported diagnostic accuracy as a primary outcome were included if a current and adequate reference standard was used.

Titles and abstracts identified in the searches were screened by one reviewer and obviously irrelevant articles excluded. A second reviewer independently screened 15% of the titles and abstracts. The second reviewer did not identify any titles or abstracts as potential inclusions

that were ultimately included in the review, but considered not relevant by the first reviewer. Potentially relevant studies were obtained in full text and independently assessed by two reviewers against the review inclusion criteria. Discrepancies were discussed and resolved by inclusion of a third reviewer.

3.4.3 Data extraction, assessment of risk of bias and data synthesis

Two reviewers independently extracted data and assessed risk of bias.

Data extraction

We extracted information on the experimental arm including:

- a) The prediction rule or diagnostic strategy tested and its role in the existing diagnostic pathway (replacement, triage or add-on). (87)
- b) Whether the strategy was assistive (e.g. provided a probability estimate or risk classification, or directive (e.g. suggested or recommended a course of action),
- c) Whether the use of the prediction tool was discretionary (e.g. the study methods stated the clinician could choose to use or not use the prediction tool) or expected (e.g. the study methods implied or stated that the prediction tools be used by clinicians) and,
- d) Whether application of the output of the CPR (when the CPR output was a suggested course of action) was discretionary (e.g. the study report stated a clinician could decide whether to follow the rule recommendation or override it) or mandatory (e.g. the recommendation of the CPR was followed in all patients).

For the control arm of the studies, we extracted the description of care as reported by the study authors. For studies describing the control arm only as 'usual practice' or similar, we noted whether the study design may have led to some modification of 'usual practice' (for example, where clinicians in the control group may have received training or information on the CPRs under evaluation).

We also assessed whether elements of the study interventions necessary for interpretation of study findings and replication in clinical practice were reported. We determined the minimum items required for reporting of the interventions through discussion and consideration of internationally accepted standards for reporting of clinical trials. (178, 179) This included a description of the diagnostic strategies tested (beyond stating the name of the test), description of the criteria used for establishing a diagnosis or treatment decision and, for studies reporting primary outcomes affected by administration of selected treatments (e.g.

patient symptoms), a description of the administered treatment. For the experimental arm, reporting of aspects of implementation (e.g. training in or exposure to the diagnostic strategy) was also assessed. The items were judged as 'reported' if any relevant information was described.

Risk of bias assessment

Risk of bias was assessed using the criteria in the Cochrane Handbook for Systematic Reviews of Interventions. (180) This domain based evaluation involved independent assessment of risk of bias due to selection, performance, detection, attrition, reporting and other. We considered the following features to judge the risk of bias for each domain; random sequence generation, allocation concealment (selection bias), blinding of participants (performance bias), blinding of outcome assessors (detection bias) incomplete outcome data (attrition bias) and selective reporting (reporting bias). Assessments of risk of bias arising from allocation concealment were based on the methods used to assign clusters of individuals (hospitals or practices), individual clinicians or individual patients to experimental or control groups. Judgements on the likelihood of detection bias were based on details about how outcomes were determined, ascertained or verified and the subjectivity of the primary outcome of the study. For trials that randomised centres, and trials that randomised individual clinicians who then recruited patients to the study, we also assessed risk of bias arising from the recruitment of patients to the study by clinicians aware of their allocation (recruitment bias). For these studies we also assessed whether the analysis had been adjusted for clustering and whether there was baseline comparability of clusters or statistical adjustment where there was imbalance. The potential for contamination was also recorded. Contamination may occur, for instance, when patients are randomised to either the intervention or control group with the clinician switching between use and no use of the prediction rule, or when clinicians within the same centre randomised to different study groups discuss their experiences. We assessed the methods as low risk, high risk or unclear risk for each domain. We resolved any disagreement through discussion or inclusion of a third reviewer.

Data synthesis

Given differences in the objectives of the included trials and clinical prediction rules applicable to each condition, we expected heterogeneity in the outcomes reported between and within

clinical conditions. To facilitate interpretation, the included studies were grouped by the clinical condition for which use of the CPR was proposed, and the results described in terms of patient outcomes (outcomes which are a direct measure of a patient's health e.g. mortality, clinical events, health related quality of life, patient symptoms and adverse events), process outcomes that are a measure of the healthcare service provided (e.g. length of stay, time to operation), clinicians' decisions (test ordering, treatment or referral decisions) or the appropriateness of their decisions, accuracy (agreement with a reference standard test), and use and implementation (use of the tool or compliance with the output of the directive CPR) of the CPR or diagnostic strategy.

We extracted and tabulated results data for all outcomes reported in the trials. For dichotomous outcomes, we presented the adjusted estimates of effect reported in the paper and calculated risk ratios and 95% confidence intervals when this was not presented. For instances where the intervention is intended to prevent an undesirable outcome (e.g. symptoms or antibiotic prescribing), an OR, HR or RR of <1 indicates the intervention is better than the control. Where the intervention is intended to promote a positive event (e.g. safe discharge) an OR, RR or HR >1 confirms treatment efficacy. Continuous outcome measures are presented as reported in the paper. For each study we identified the primary outcome. The primary outcome was considered to be either the outcome stated by the study as being the primary outcome, or the outcome for which a power calculation was conducted. In the absence of these, the primary outcome was considered to be the outcome mentioned in the study objective or reported first in the results section. Statistical analysis was carried out using the Cochrane Collaborations' Review Manager (Version 5.3) software. When five or more studies for a clinical condition assessed the same outcome, in the same manner, regardless of the specific prediction tool or diagnostic strategy, we obtained a pooled estimate of effect using the general inverse variance method. We chose the risk ratio, a relative measure of effect as the summary statistic for its ease of interpretation. (181) We considered clinical heterogeneity sufficient to expect that the underlying treatment effects differed between trials. Consequently, we used random effects meta-analysis to produce an overall summary of the average treatment effect across the included studies. For studies including two experimental arms (e.g. CPR and CPR plus RADT), (182, 183) we included data only from the CPR alone arm. For the one study including two control arms (e.g. clinical judgment with no diagnostic aid or with a standardised data collection form), (184) we included data only from the clinical judgment with no diagnostic aid arm. To pool individual and cluster randomised trials in the same model, adjustment for clustering was conducted. Adjustment involved

reducing the size of the cluster trials to the effective sample size by dividing the sample size by the design effect, where the design effect is equal to $1 + (m-1) \times \text{intraclass correlation coefficient (ICC)}$ and m is the average cluster size. (180) To calculate the design effect for studies of sore throat we used the intraclass correlation coefficient reported in a duplicate publication of one of the other sore throat studies. (8) For appendicitis studies, we used the median ICC reported for implementation research studies reporting process variables (ICC 0.063), and undertook sensitivity analysis using the extremes of the interquartile range. (185)

3.5 Results

3.5.1 Study selection

Of 10,351 titles and abstracts screened, 166 were obtained in full text and 25 studies were included in the review (Figure 3.1). (60, 171, 182-184, 186-205) The 141 excluded studies are presented in Appendix B with reasons for their exclusion. One study (194) evaluated two different prediction rules for different clinical conditions. Two studies (182, 183) compared two different diagnostic strategies incorporating the same CPR (e.g. Centor score alone and Centor score with rapid antigen detection testing), and one study compared the experimental arm to a control arm with no diagnostic aids and to clinicians using a standardised data collection form, (184) so there are 29 comparisons between a group exposed to a diagnostic CPR alone or as part of a diagnostic strategy, and a control group with no exposure to a diagnostic CPR.

3.5.2 Trial characteristics

Characteristics of the 25 included trials grouped by the clinical condition for which the CPR was developed, are presented in Table 3.1.

The prospective role of the CPR or diagnostic strategy incorporating a CPR was usually not specified in the study report, but was determined based on aspects of study design or descriptions provided in the study introduction or objectives. Nine of the 25 included studies were considered to have evaluated a diagnostic CPR or strategy designed to replace the existing approach, (60, 182, 186, 190, 195, 196, 201, 202, 205) while 12 assessed the impact of adding a CPR to the usual diagnostic pathway in order to evaluate the benefit of extra information to diagnostic decision making. (171, 183, 187, 189, 192-194, 198-200, 203, 204) One study with 3 intervention arms evaluated a CPR as both a replacement and add-on test. (184) Three of the 25 studies evaluated CPRs as a triage test. In these studies, the CPR was used before the existing test to determine which patients undergo the existing test. (188, 191, 197)

In 18 of 25 studies, the CPR was introduced as a stand-alone tool (171, 184, 186, 188-196, 198-200, 202, 204, 205) and in 5 studies, (60, 187, 197, 201, 203) the CPR was part of a diagnostic pathway with other tests (e.g. a pathway including a CPR, electrocardiogram and cardiac troponin tests). Two studies evaluated both the introduction of a CPR alone and a CPR in combination with another diagnostic test. (182, 183) The diagnostic CPRs and strategies tested were directive in 23 of the 25 studies, making a recommendation regarding treatment or disposition in 8 studies, (60, 183, 186, 189, 192, 193, 201, 205) further diagnostic testing in 8 studies, (171, 188, 191, 197-200, 204) and both further testing and treatment in 7 studies. (182, 187, 190, 194-196, 202)

In most studies (16/25), the control group intervention was variably described as 'clinicians' assessment' or 'usual care'. (171, 183, 187-194, 198-200, 202-204) This ranged from control groups in which clinicians were explicitly asked not to change their usual practice, and control groups where clinicians received information on the CPRs being tested in the intervention arm, to control groups where clinicians' actions were expected to be based on local care guidelines. In 2 studies, clinicians in the control group were required to use a standard data collection form (195, 196) and in 4 studies, care provided by the control group was a specific management pathway (e.g. a chest pain clinic protocol or delayed antibiotic treatment strategy). (60, 182, 201, 205) In 2 studies, the control group intervention was a single imaging test, (186, 197) and in 1 study with 3 intervention arms, the control interventions were clinicians' usual care and clinicians using a structured data collection form. (184)

Patient outcomes were considered the primary outcome in 3 (182, 190, 197) of the 25 studies, and process of care outcomes (e.g. length of stay) were the primary outcome in 4. (187-189, 198) A clinicians' decision was the primary outcome in 11 studies, (60, 171, 183, 191, 194, 195, 199, 200, 202, 204, 205) and the appropriateness of the decision the primary outcome in 3. (196, 201, 203) Accuracy of a diagnosis or decision was the primary outcome in 4 studies. (184, 186, 192, 193) The types of primary and secondary outcomes reported in the included studies are shown in Appendix B.

Figure 3.1. Study flow diagram

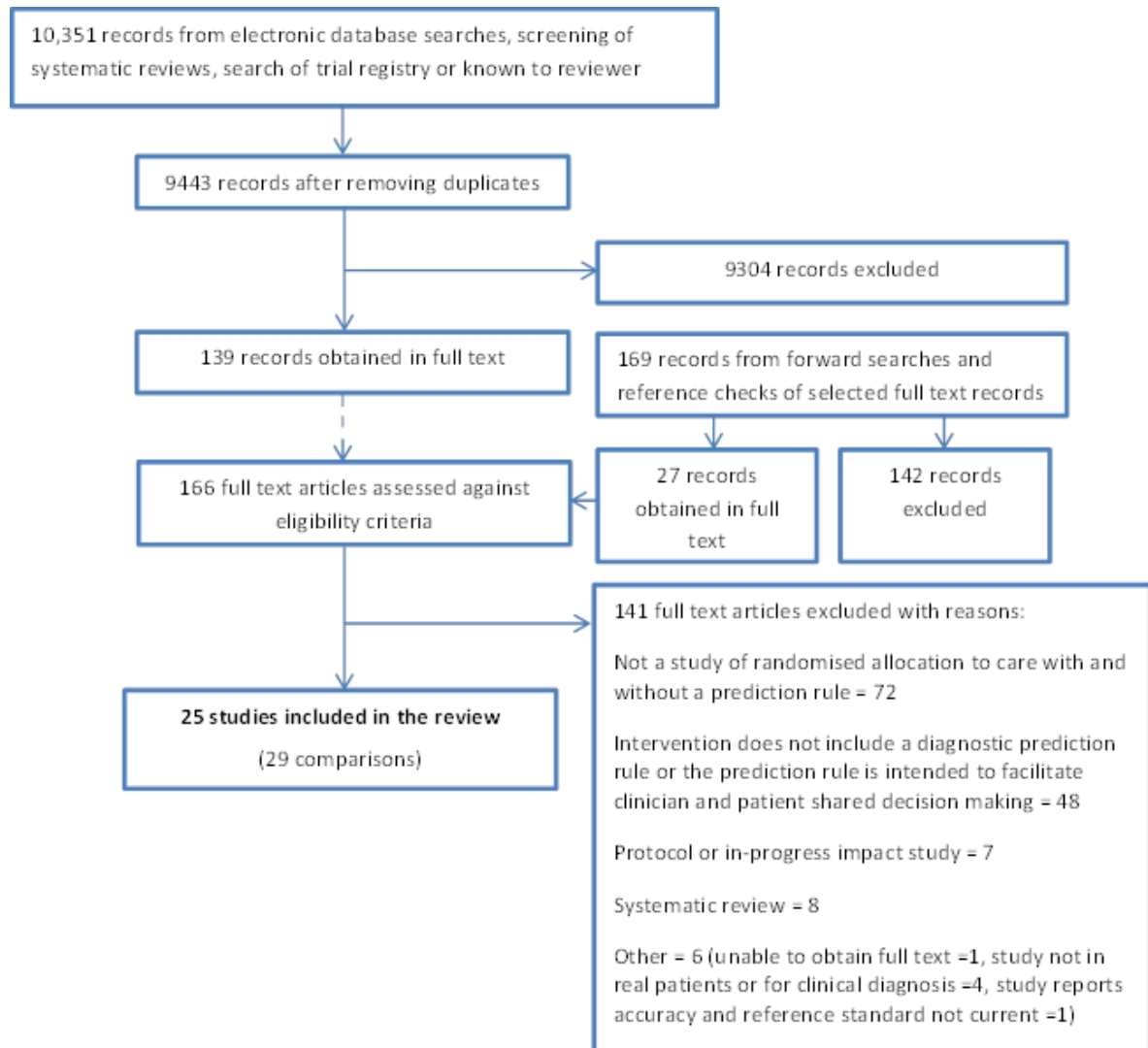


Table 3.1. Characteristics of the included studies by clinical condition

Study/location/Setting	Study arm	Diagnostic strategies tested (output format of the CPR or diagnostic strategy)	Interventions	Proposed role of the CPR or experimental diagnostic strategy	Use of the CPR or experimental diagnostic strategy	Application of the output of the CPR or experimental diagnostic strategy	Primary outcome of the study*
Group A streptococcus infection of the throat							
Worrall et al., 2007 (183)/Canada/PC	Experimental	Clinicians' usual practice + Centor Score (D)		Add-on	Expected	Discretionary	Clinicians' decision
	Experimental	Clinicians' usual practice + Centor Score + RADT (D)					
	Control	Clinicians' usual practice					
McIsaac & Goel 1998 (195)/Canada/PC	Experimental	Clinician + Centor Score (D)		Replacement	Expected	Discretionary	Clinicians' decision
	Control	Clinician + structured clinical checklist					
	Experimental	Clinician + Modified Centor Score (D)					
McIsaac et al., 2002 (196)/Canada/PC	Experimental	Clinician + structured clinical checklist		Replacement	Expected	Discretionary	Clinicians' decision
	Control	Clinician + structured clinical checklist					
	Experimental	Clinicians' usual care + Walsh Rule (D)					
McGinn et al., 2013 (194)/USA/PCT	Experimental	Clinicians' usual care†		Add-on	Discretionary	Discretionary	Clinicians' decision
	Control	Clinicians' usual care†					
	Experimental	Clinician + FeverPAIN Score (D)					
Little et al., 2013 (182)/United Kingdom/PC	Experimental	Clinician + FeverPAIN Score + RADT (D)		Replacement	Expected	Discretionary	Patient outcome
	Experimental	Clinician + strategy of delayed antibiotics					
	Control						
Acute appendicitis							
Douglas et al., 2000 (187)/Australia/SU	Experimental	Clinicians' clinical diagnosis + Alvarado Score + US (D)		Add-on	Expected	Discretionary§	Process of care
	Control	Clinicians' clinical diagnosis†					
	Experimental	Clinicians' assessment + Alvarado Score (D)					
Farahnak et al., 2007 (189)/Iran/ED	Experimental	Clinicians' assessment		Add-on	Expected	Discretionary	Process of care
	Control	Clinicians' assessment					
	Experimental	Clinicians' assessment + Lintula Score (D)					
Lintula et al., 2010 (193)/Finland/ED	Experimental	Clinician's assessment†		Add-on	Expected	Discretionary	Accuracy
	Control	Clinicians' assessment + Lintula Score (D)					
	Experimental	Clinicians' assessment + Lintula Score (D)					
Lintula et al., 2009 (192)/Finland/ED	Experimental	Clinicians' assessment†		Add-on	Expected	Discretionary	Accuracy
	Control	Clinicians' assessment†					
	Experimental	Clinicians' assessment + Leeds Decision Support System (A)					
Wellwood et al., 1992 (184)/United Kingdom/ED	Experimental	Clinician with no diagnostic aid		Add-on/ Replacement	Expected	NA	Accuracy
	Control	Clinician + structured data collection form					
	Control						

Study/location/Setting	Study arm	Diagnostic strategies tested (output format of the CPR or diagnostic strategy)	Proposed role of the CPR or experimental diagnostic strategy	Use of the CPR or experimental diagnostic strategy	Application of the output of the CPR or experimental diagnostic strategy	Primary outcome of the study*
Serious bacterial infection in children with fever						
Roukema et al., 2008 (198)/The Netherlands/ED	Experimental	Clinicians' assessment + Prediction rules of Bleeker (D)	Add-on	Expected	Discretionary	Process of care
	Control	Clinicians' assessment				
	Experimental	Clinician + LAB Score (procalcitonin, CRP, Urinary dipstick) (D) blind to WBC count and differential	Replacement	Expected	Discretionary	Clinicians' decision
Lacroix et al., 2014 (205)/Switzerland/ED	Control	Clinician + WBC count , band count and CRP, blind to procalcitonin and LAB Score				
de Vos-Kerkhof et al., 2015 (204)/The Netherlands/ED	Experimental	Clinicians' usual care + Rule of Nijman (D)	Add-on	Expected	Discretionary	Clinicians' decision
	Control	Clinicians' usual care				
Ankle/foot fracture						
Auleley et al., 1997 (171)/France/ED	Experimental	Clinicians' usual practice + Ottawa Ankle Rules (D)	Add-on	Discretionary	Discretionary	Clinicians' decision
	Control	Clinicians' usual practice‡				
Fan et al., 2006 (188)/Canada/ED	Experimental	Ottawa Ankle Rules (D): if positive x-ray, if negative clinical assessment	Triage	Expected	Mandatory	Process of care
	Control	Standard departmental care				
Acute coronary syndrome						
Than et al., 2014 (201)/New Zealand/ED	Experimental	Accelerated diagnostic pathway: TIMI Score, ECG + troponin at presentation and 2 hours after symptom onset (D)	Replacement	Expected	Discretionary	Clinicians' decision
	Control	Standard-care chest pain pathway: initial ECG + troponin at presentation and 6-12 hours after symptom onset				
Sanchis et al., 2010 (60)/Spain/ED	Experimental	Sanchis risk score + NT-proBNP (D)	Replacement	Expected	Discretionary	Process of care
	Control	Chest pain unit protocol with early exercise testing				

Table 3.1. Continued

Study/location/Setting	Study arm	Diagnostic strategies tested Interventions output format of the CPR or diagnostic strategy	Proposed role of the CPR or experimental diagnostic strategy	Use of the CPR or experimental diagnostic strategy	Application of the output of the CPR or experimental diagnostic strategy	Primary outcome of the study*
Bacterial pneumonia						
Torres et al., 2014 (202)/Argentina/OC	Experimental Control	Bacterial pneumonia score (D) Standard management based on institutional guidelines	Replacement	Expected	Mandatory	Clinicians' decision
Pneumonia						
McGinn et al., 2013 (194)/USA/ED†	Experimental Control	Clinicians' usual care + Walsh Rule (D) Clinicians' usual care‡	Add-on	Discretionary	Discretionary	Clinicians' decision
Joint or bone injuries of the extremities in children						
Klassen et al., 1993 (191)/Canada/ED	Experimental Control	Brand protocol (D) : if positive x-ray, if negative clinical assessment Standard care	Triage	Expected	Mandatory	Clinicians' decision
Suspicious pigmented skin lesion						
Walter et al., 2012 (203)/United Kingdom/PC	Experimental Control	Best practice: history, naked eye examination, seven point checklist + Primary care scoring algorithm + SIAscopy scanner (A) Best practice: history, naked eye examination, seven-point checklist	Add-on	Expected	Discretionary	Clinicians' decision
Pulmonary embolism						
Rodger et al., 2006 (197)/Canada/NMD	Experimental Control	Bedside tests (D): Wells PE score , D-dimer, AVDSf – if 22 tests positive VQ scan Initial VQ scan blind to bedside tests	Triage	Expected	Mandatory	Patient outcome

Table 3.1. Continued

Study/location/Setting	Study arm	Diagnostic strategies tested Interventions (output format of the CPR or diagnostic strategy)	Proposed role of the CPR or experimental diagnostic strategy	Use of the CPR or experimental diagnostic strategy	Application of the output of the CPR or experimental diagnostic strategy	Primary outcome of the study*
Gastro-oesophageal reflux disease						
Horowitz et al., 2007 (190)/Israel/PC	Experimental Control	Algorithm (D): alarm symptom assessment – if positive gastroscopy, if negative Gerd score- if negative C-Urea Breath test Clinicians' discretion	Replacement	Expected	Mandatory	Patient outcome
Acute small bowel obstruction						
Bogusevicius et al., 2002 (186)/Lithuania/SU	Experimental Control	Rule of Bogusevicius (D) Contrast radiography	Replacement	Expected	Mandatory	Accuracy
Clinically important brain injury						
Stiell et al., 2010 (200)/Canada/ED	Experimental Control	Clinicians' usual practice + Canadian CT Head Rule (D) Clinicians' usual practice	Add-on	Expected	Discretionary	Clinicians' decision
Cervical spine fracture						
Stiell et al., 2009 (199)/Canada/ED	Experimental Control	Clinicians' usual practice + Canadian C-Spine Rule (D) Clinicians' usual practice	Add-on	Expected	Discretionary	Clinicians' decision

CPR – clinical prediction rule; (D) Directive output format i.e. Suggests a course of action; (A) Assistive output format i.e. Provides a probability without suggesting course of action; RADT- rapid antigen detection test; US – ultrasound; CRP – C-reactive protein; AVDSF- alveolar dead-space fraction; VQ- ventilation perfusion scan; PC – primary care; SU – surgical unit; ED – Emergency department; OC – Outpatient clinic; NMD – nuclear medicine department; ECG – electrocardiogram

* The primary study outcome was the outcome stated by the study as being the primary outcome, or the outcome for which a power calculation was conducted. In absence of these, the primary outcome was considered to be the outcome mentioned in the study objective or reported first in the results section. Patient outcomes are direct measures of patients health e.g. symptoms, clinical events. Process of care outcomes are measures of the healthcare provided e.g. length of stay, time to operation; †this study evaluated CPRs for different clinical conditions; ‡ the diagnostic strategy may be modified by the provision of information related to the CPRs being tested; § application mandatory only for certain patients; || different rules for self-referred and clinician referred patients

3.5.3 Risk of bias

The majority of studies included in the review (19/25 (76%)) were judged to be at unclear or high risk of bias on 3 or more domains (including performance bias). Concealment of allocation, one of the key domains in the assessment of bias, was reported in insufficient detail to enable accurate judgment in over half of the included trials (16/25 (64%)). Due to the nature of the intervention, in most studies included in this review, clinicians would have been aware of group allocation. We judged the impact of this on risk of performance bias to be high in the majority (22/25 (87%)) of the included trials and unclear in 1 study. In this study the interventions were very similar and the study stated that clinicians were not aware of the alternate interventions. The risk of performance bias was considered to be low in 3 studies where the CPR was used as a triage test. In 2 of these studies, triage nurses used the CPR and care was provided by clinicians unaware of whether the patient had entered through the control protocol or was negative according to the experimental protocol. (188, 191) In the other study, the clinicians using the CPR and providing care were different, and sham procedures were also applied. (197) The likelihood of detection bias was judged to be high or unclear in 19 of 25 (74%) studies, and low in 6. The potential for selective reporting bias could not be determined in the majority of studies (18/25 (72%)). Contamination was judged to be possible in 3 of the 25 studies and an unclear risk in 3. The majority (8/10) of studies randomising centres, or individual clinicians' who recruited patients, adjusted for clustering using appropriate methods. Details of the risk of bias assessment for each trial are shown in Table 3.2.

Table 3.2. Risk of bias in the included studies

Domain of bias	Selection bias		Performance bias	Detection bias	Attrition bias	Reporting bias	Other bias		
	Random sequence generation	Allocation concealment	Blinding of participants/personnel	Blinding of outcome assessment	Incomplete outcome data	Selective reporting	Recruitment bias*	Baseline imbalance*	Incorrect analysis*
Sore throat									
Worrall et al 2007 (183)	Low	Unclear	High	Unclear	High	Unclear	Unclear	Unclear	High
McIsaac & Goel 1998 (195)	Unclear	Unclear	Unclear	Unclear	High	Unclear	-	-	Low
McIsaac et al 2002 (196)	Unclear	Unclear	High	Unclear	High	Unclear	High	Low	Low
McGinn et al 2013† (194)	Low	Unclear	High	Unclear	Low	Low	Low	Low	Unclear
Little et al 2013 (182)	Low	Low	High	Unclear	Low	Low	-	-	High
Acute appendicitis									
Douglas et al 2000 (187)	Low	High	High	Unclear	Low	Unclear	-	-	Low
Farahnak et al 2007 (189)	Low	Unclear	High	Unclear	Low	Unclear	-	-	Low
Lintula et al 2010 (193)	Unclear	Unclear	High	Unclear	Low	Unclear	-	-	High
Lintula et al 2009 (192)	Unclear	Low	High	Unclear	Low	Unclear	-	-	High
Wellwood et al 1992 (184)	Low	Unclear	High	Unclear	Low	Unclear	Low	Unclear	Unclear
Serious infection in children with fever									
Roukema et al 2008 (198)	Low	Unclear	High	Low	High	Unclear	-	-	Low
Lacroix et al 2014 (205)	Low	Unclear	High	Unclear	Low	Low	-	-	Low
de Vos-Kerkhof et al 2015 (204)	Unclear	Unclear	High	Unclear	Low	Unclear	-	-	Low
Ankle or mid-foot injury									
Auleley et al 1997 (171)	Unclear	Unclear	High	Unclear	Low	Unclear	Low	Unclear	Low
Fan et al 2006 (188)	Unclear	Low	Low	Low	High	Unclear	-	-	Low
Possible cardiac chest pain									
Than et al 2014 (201)	Low	Low	High	Low	Low	Low	-	-	Low
Sanchis et al 2010 (60)	Low	Low	High	Unclear	Low	Low	-	-	Low
Single studies of different clinical conditions									
Torres et al 2014 (202)	Unclear	Unclear	High	Low	Low	Unclear	-	-	Low
McGinn et al 2013† (194)	Low	Unclear	High	Unclear	Low	Low	Low	Low	Unclear
Klassen et al 1993 (191)	Low	Unclear	Low	Unclear	Low	Unclear	-	-	Low
Walter et al 2012 (203)	Low	Low	High	High	Low	Low	-	-	Low
Rodger et al 2006 (197)	Unclear	Low	Low	Low	Low	Unclear	-	-	Low
Horowitz et al 2007 (190)	Unclear	Unclear	High	High	Low	Unclear	Unclear	Unclear	High
Bogusevicius et al 2002 (186)	Unclear	Low	High	Unclear	Low	Unclear	-	-	Low
Stiell et al 2010 (200)	Unclear	Unclear	High	Unclear	Low	Unclear	Low	Low	Low
Stiell et al 2009 (199)	Low	Unclear	High	Unclear	Low	Low	Low	Low	Low

* For trials randomising centres or clinics or individual clinicians who then recruit participants to the study; † this study evaluated two prediction rules for different clinical conditions

3.5.4 Effects of diagnostic strategies incorporating diagnostic clinical prediction rules

The estimated effects of exposure to a diagnostic strategy incorporating a diagnostic CPR are presented in Tables 3.3 to 3.8 and Figures 3.2 and 3.3.

Studies of Group A streptococcus throat infection (Table 3.3 and Figure 3.2)

5 studies evaluated 3 different CPRs (Walsh, Centor and FeverPAIN score) and a modified version of one (Modified Centor Score). (182, 183, 194-196) All CPRs evaluated were directive (i.e. provided management recommendations) and in all 5 studies, application of the output of the CPR or diagnostic strategy was discretionary (i.e. the clinician could follow, or not follow the recommendations of the CPR). Four of the five studies were judged to be at high or unclear risk of bias on 3 or more of the 6 key domains.

Clinical outcomes

In one trial reporting patient reported symptoms (primary study outcome) and adverse effects (182), there were greater improvements in symptom severity among patients randomised to the FeverPAIN score compared to the control arm with a strategy of delayed antibiotics (Mean difference adjusted for baseline symptom severity and fever -0.33 on a score of 0-6, 95% CI -0.64 to -0.02 $p=0.04$). The combination of FeverPAIN score and rapid antigen detection test (RADT) had similar effects on symptoms as the score alone (approximately 1 person in 3 rated score throat and difficulty swallowing as slight rather than moderately bad), however the effect of the combination of score and rapid antigen detection test (RADT) was not clinically or statistically significantly different to the control arm (0.30 95% CI -0.61 to 0.004 $p=0.05$). Symptom resolution was faster among patients randomised to the FeverPAIN score (HR adjusted for baseline symptom severity and fever 1.30 95% CI 1.03 to 1.63; Median duration 5 days (IQR 3-7) in the control arm and 4 days (IQR 2-6) in the score only arm). There were no differences between the study groups in the proportion returning to the clinic with sore throat within a 1 month period or the occurrence of suppurative complications (none occurred during the trial).

Clinicians' decisions

All 5 trials reported clinicians' decisions to prescribe antibiotics (this was the primary outcome in 3 trials). (182, 183, 194-196) In pooled analysis, clinical prediction rules reduced antibiotic prescriptions compared to care provided without a CPR (pooled RR 0.86 95% CI 0.75 to 0.99) (Figure 2). Two of these 5 trials contained 2 experimental arms; 1 arm evaluated the CPR alone and the second arm a strategy of CPR and rapid antigen diagnostic testing (RADT). In one of

these studies, the combination of CPR and RADT decreased antibiotic prescriptions compared to usual clinical judgment but use of the CPR alone did not (% of visits where antibiotics prescribed; 58.2% usual practice, 55.3% CPR alone $p=0.58$, 38.2% CPR plus RADT $p<0.00$). (183) In the other study, both strategies reduced antibiotic prescribing compared to a strategy of delayed antibiotic prescribing (adjusted RR CPR alone 0.71 95%CI 0.50 to 0.95, CPR plus RADT 0.73 95%CI 0.52 to 0.98). (182)

Two trials reported the appropriateness of antibiotic prescribing. The first (196) reported no difference between the study arms in unnecessary antibiotic prescriptions defined as a prescription for an antibiotic in a patient with a throat culture (which was performed in all patients) negative for Group A streptococcus (OR adjusted 0.76 95%CI 0.42 to 1.40). In the second study (195) appropriateness of prescribing was determined by assessing whether the prescribing decisions of the experimental group were more like the management recommendations of the CPR than the control group, and by comparing the proportions of patients prescribed antibiotics at different levels of probability of GAS infection. In the experimental group, antibiotic prescribing and throat culture use corresponded more closely to suggested management recommendations, and a greater reduction in prescribing was observed for patients with low probability of GAS infection.

In one trial evaluating the effect of the Walsh rule, clinicians in the experimental group were significantly less likely to order rapid streptococcal tests (RR adjusted 0.75 95%CI 0.58 to 0.97), but not pharyngitis throat cultures (RR adjusted 0.54 95% CI 0.18 to 1.64). (194)

Use and application outcomes

In the one trial where use of the CPR was discretionary, (194) clinicians opened the computerised CPR tool for 74.3% (278/374) of eligible patients and opened the risk score calculator embedded within the tool for 66.6% (249/374) of eligible patients.

Table 3.3. Results of studies of Group A Streptococcus throat infection by outcome

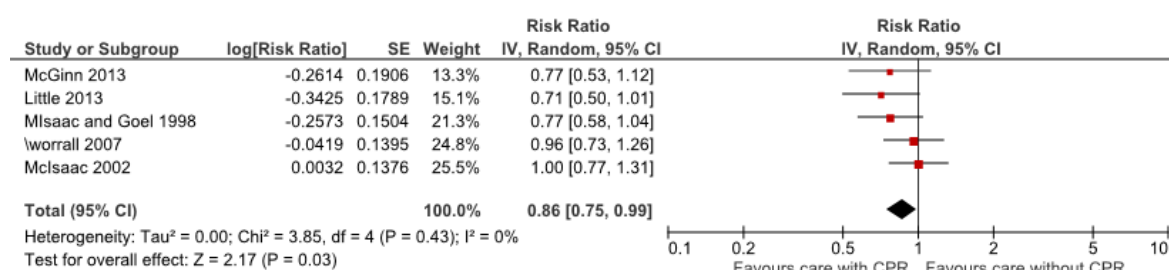
Outcomes	Number of trials/ N of trial or comparison	Result
Patient outcomes		
Symptom severity (6 point scale)	1 (Little et al., (182))*/336 1 (Little et al., (182))*/334	CPR only MD [†] -0.33 (95% CI -0.64 to -0.02) CPR + RADT MD [†] -0.30 (95% CI -0.61 to 0.00)
Resolution of symptoms rated moderately bad or worse		CPR only HR [‡] 1.30 (95% CI 1.03 to 1.63) CPR + RADT HR [‡] 1.11 (95% CI 0.88 to 1.40)
Return within 1 month with sore throat		CPR only RR [‡] 0.91 (95% CI 0.47 to 1.72) CPR + RADT RR [‡] 0.74 (95% CI 0.36 to 1.47)
Suppurative complications		no suppurative complications during the trial
Clinicians' decisions/appropriateness of clinicians' decisions		
Antibiotic prescriptions	5 (Little et al., (182), McIsaac et al., (196), Worrall et al., (183)*, McIsaac & Goel (195)*, McGinn et al., (194)*)	CPR only pooled RR 0.86 (95%CI 0.75 to 0.99)
	1 (Little et al., (182))/334 1 (Worrall et al.,(183))/243	CPR + RADT RR [‡] 0.73 (95%CI 0.52 to 0.98) CPR + RADT RR 0.66 (95%CI 0.50 to 0.87)
Unnecessary antibiotic prescriptions [§]	1 (McIsaac et al.,(196)*)/621	OR 0.76 [‡] (95%CI 0.42 to 1.40)
Antibiotic prescription by CPR recommendation	1 (McIsaac & Goel (195))/396	
No swab, no script		OR 0.30 (95%CI 0.05 to 1.18)
Swab, no script		OR 0.55 (95%CI 0.26 to 1.15)
Script if early or swab		OR 0.71 (95%CI 0.01 to 59.77)
Throat swab by CPR recommendation	1 (McIsaac & Goel(195))/396	
No swab, no script		OR 0.80 (95%CI 0.44 to 1.44)
Swab, no script		OR 2.44 (95%CI 0.84 to 7.67)
Script if early or swab		OR 1.73 (95%CI 0.289 to 12.70)
Rapid streptococcal test orders	1 (McGinn et al.,(194))/598	RR [¶] 0.75 (95%CI 0.58 to 0.97)
Pharyngitis throat culture orders	1 (McGinn et al.,(194))/598	RR [¶] 0.54 (95%CI 0.18 to 1.64)
Use and application outcomes **		
Decision support tool opened	1 (McGinn et al.,(194))/598	74.3%
CPR calculator opened		66.6%

CPR – clinical prediction rule; MD – mean difference; RADT – rapid antigen detection test; HR – hazard ratio; OR – odds ratio; RR – risk ratio

*primary outcome of the trial; †adjusted for baseline severity and fever; ‡Adjusted for baseline severity, fever and previous antibiotic use; §prescription of an antibiotic when throat culture negative; ¶ Adjusted for patient and physician characteristics;

¶Adjusted for age; **only relevant to the intervention arm of studies where use of the CPR was discretionary

Figure 3.2. Meta-analysis of Group A Streptococcus throat infection studies for the outcome antibiotic prescriptions



Studies of acute appendicitis (Table 3. 4 and Figure 3.3)

Five studies evaluated 3 different CPRs (Alvarado score, Lintula score and the Leeds Decision Support System). (184, 187, 189, 192, 193) The Leeds decision support system was assistive, providing only an estimate of the probability of appendicitis without recommending a course of action, and application of the management recommendations of the Alvarado and Lintula scores was discretionary. All 5 studies were judged to be at high or unclear risk of bias arising from lack of blinding of care providers and outcome assessors.

Clinical outcomes

Perforated appendix rates did not significantly differ between the experimental group and a control group providing care without a diagnostic aid (RR 0.47 95% CI 0.19 to 1.15) and a control group where clinicians used a standard data collection form (RR 0.81 95% CI 0.31 to 2.16). (170, 184)

Process of care outcomes

The results of two studies providing data on the effect of CPRs on duration of hospitalisation among patients with suspected appendicitis are conflicting. One small study of the Alvarado score reported significantly shorter duration of hospitalisation in the intervention group (Median 37.00 hrs vs 60.40 hrs $p=0.03$), (189) while the other study reported no difference in mean duration of hospital stay between a diagnostic protocol incorporating the Alvarado score and graded compression ultrasound, and the control group (Mean 53.4 hrs vs 54.5 hrs $p=0.84$). (187)

The effect of CPRs on admission rates was conflicting. Admission rates were reduced in one study compared to clinical judgment without a diagnostic aid (RR 0.90 95% CI 0.82 to 0.99). (184) CPRs did not reduce admission rates in 2 studies compared to a control group where clinicians used a standard data collection form (RR 1.0 95%CI 0.91 to 1.12)(184) and usual clinical assessment(RR 0.72 95%CI 0.49 to 1.05). (189) There was an increase in delayed treatment in association with perforation in the experimental arm of two trials, but the difference was not statistically significant (RR 3.0 95% CI 0.13 to 69.7, (189) and RR 2.22 95%CI 0.44 to 11.26). (187)

In 1 small trial ($n=42$) reporting time to surgery (from randomisation to skin preparation), time to surgery was significantly shorter in the experimental group (Median 2.05 hrs vs 8.35 hrs $p=0.03$). (189)

One trial evaluated time to therapeutic operation (an operation was considered therapeutic if the disease found seemed to be the cause for the patients pain, and surgery was considered the appropriate treatment for the disease), and found that patients in the experimental group (diagnostic protocol incorporating Alvarado score and grade compression ultrasound) who underwent therapeutic operation had significantly shorter time to operation than patients in the control group (mean 7.0 hrs vs 10.2 hrs $p=0.016$). (187)

Clinicians' decisions/appropriateness of clinicians' decisions

In pooled analysis of five trials, diagnostic strategies incorporating CPRs reduced unnecessary surgeries compared to usual clinical assessment, but this was not statistically significant (pooled RR 0.68 95% CI 0.43 to 1.08). (184, 187, 189, 192, 193) The direction of effect was consistently in favour of the experimental arms of the trials, though the risk ratios varied widely (Figure 3). An ICC obtained from an analysis of implementation research studies reporting process outcomes (ICC 0.063) was used to adjust for clustering in the one cluster randomised trial included in this analysis. In sensitivity analysis, using the lower extreme of the ICC interquartile range, in which more weight is given to the study in the meta-analysis similar to that applied when unadjusted data are used, the confidence intervals were narrower and the effect significant (pooled RR 0.64 95% CI 0.41 to 0.98). The results of the studies in the main and sensitivity analysis are homogenous.

Accuracy

3 trials report indices of accuracy. (184, 192, 193) The reference standard was a diagnosis obtained at surgery or at discharge in 1 trial (184) or at 1 month follow-up in 2 trials. (192, 193) In these two trials, accuracy of the presumptive diagnosis was compared at initial and final (> 3 hours after initial) examination.

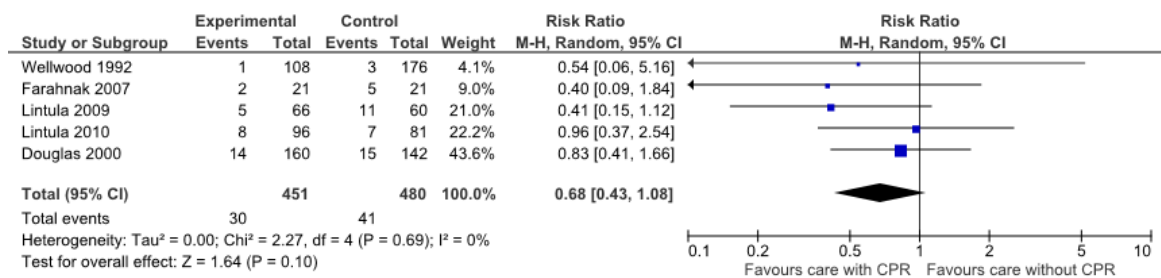
In one trial in children, there was no difference in sensitivity and specificity of the initial examination between the experimental and control groups (sensitivity 83% vs 85 p value not significant and specificity 69% vs 52 % p value not significant respectively) (192). In one trial in adults there was no difference in sensitivity (87% vs 89% p value not significant) but specificity of initial clinical assessment was higher (59% vs 80% $p=0.03$). (193) In the third trial, sensitivity and specificity in the experimental arm was higher than clinical judgment with no diagnostic aids (sensitivity 48% vs 28% $p=0.01$ and specificity 98% vs 96% $p=0.04$) and clinical judgment with standardised data collection (sensitivity 48% vs 42% $p=0.48$ and specificity 98% vs 96% $p=0.01$). (184)

Table 3.4. Results of studies of acute appendicitis by outcome

Outcomes	Number of trials/N of trial or comparison	Result	
		Intervention	Control
Patient outcomes			
Perforated appendix rate	1 (Wellwood et al., (170, 184)*)/4194	CPR vs judgment with no diagnostic aid RR 0.47 (95%CI 0.19 to 1.15) CPR vs judgment with SDC form RR 0.81 (95%CI 0.31 to 2.16)	
Process outcomes			
Time to therapeutic operation (mean hrs)	1 (Douglas et al., (187)†)/302	7.0 (95% CI 5.9 to 8.1)	10.2 (95%CI 7.9 to 13) p=0.01
Time to surgery (median hrs)	1 (Farahnak et al., (189)†)/42	2.05	8.35 p=0.03 [‡]
Duration of hospitalisation (mean hrs)	1 (Douglas et al., (187))/302	53.4 (95% CI 47 to 60)	54.5 (95%CI 46 to 63) p=0.84 [‡]
Duration of hospitalisation (median hrs)	1 (Farahnak et al., (189))/42	37.00	60.40 p=0.03 [‡]
Clinicians' decisions/appropriateness of clinicians' decisions			
Nontherapeutic operations	5 (184, 187, 189, 192, 193)/931	Pooled RR 0.68 (95%CI 0.43 to 1.08)	
Admissions	2 (Farahnak et al., (189))/42 (Wellwood et al., (184))/2596 (Wellwood et al., (184))/2584	RR 0.72 (95%CI 0.49 to 1.05) CPR vs judgment with no diagnostic aid RR 0.90 (95%CI 0.82 to 0.99) CPR vs judgment with SDC form RR 1.0 (95%CI 0.91 to 1.12)	
Delayed treatment in association with perforation§	2 (Farahnak et al., (189))/42 (Douglas et al., (187))/302	RR 3.0 (95%CI 0.13 to 69.7) RR 2.22 (95%CI 0.44 to 11.26)	
Accuracy			
Sensitivity of initial examination	3 (Lintula et al.,(192)†)/126 (Lintula et al.,(193)†)/177 (Wellwood et al.,(184)†)/2596 (Wellwood et al.,(184))/2584	83% 87% 48%	85% p=NS [‡] 89% p=NS [‡] 28% judgment with no diagnostic aid p=0.01 [‡] 42% judgment with SDC form p=0.48 [‡]
Specificity of initial examination	3 (Lintula et al., (192))/126 (Lintula et al., (193))/177 (Wellwood et al., (184))/2596 (Wellwood et al., (184))/2584	69% 59% 98%	52% p=NS [‡] 80% p=0.03 [‡] 96% judgment with no diagnostic aid p=0.04 [‡] 96% judgment with SDC form P=0.01 [‡]
Sensitivity of final examination [¶]	2 (Lintula et al., (192))/126 (Lintula et al., (193))/177	100% 87%	96% p=NS [‡] 100% p=0.02 [‡]
Specificity of final examination [¶]	2 (Lintula et al., (192))/126 (Lintula et al., (193))/177	88% 98%	67% p=0.03 [‡] 84% p=0.03 [‡]

CPR – clinical prediction rule; NS – not significant; SDC – structured data collection

*Data from this outcome are from the original study report and systematic review by the same author; †primary outcome; ‡p value for difference between intervention and control groups; §patients with perforation where surgery not started within 10 hours of randomisation; ||>3 hours after initial examination

Figure 3.3. Meta-analysis of acute appendicitis studies for the outcome unnecessary appendectomies**Studies of serious bacterial infection in children with fever (Table 3.5)**

Three studies evaluated 3 different directive CPRs. (198, 204, 205) There was a non-statistically significant increased length of stay in the emergency department with diagnostic strategies incorporating a CPR in the 2 studies reporting this outcome (138 v 123 median minutes $p=0.16$, and 117 vs 114 median minutes $p=0.05$ in experimental and control groups of both studies respectively). (198, 204) A CPR comprised of 3 laboratory tests (procalcitonin, C-reactive protein and urinary dipstick) did not reduce antibiotic prescribing compared to the judgment of clinicians with access to the results and associated treatment recommendation of the white blood cell count, band count and C-reactive protein tests (prescriptions for 41% and 42% of experimental and control groups respectively $p=0.88$). (205) A CPR also had no effect on the use of diagnostic tests when serious bacterial infection was and was not present as judged by a reference standard (60% vs 57% of children without pneumonia in the experimental and control groups respectively received chest x-ray $p>0.05$, and 67% vs 53% of children without urinary tract infection had urine culture performed $p>0.05$). (204) This CPR resulted in a significant increase in the number of urine dipstick tests, a significant reduction in the number of full blood count tests, and a non-significant decrease in the number of diagnostic tests overall compared to the control group.

Studies of ankle or mid-foot fracture (Table 3.6)

Two studies evaluated the impact of the Ottawa Ankle Rules (OARs). (171, 188) In 1 trial, the OARs were used as a triage test. (188) In this trial, standard departmental care was compared to a pathway in which the OARs were applied at presentation: if positive, the patient was x-rayed, and if negative, the patient underwent usual clinical assessment. In the second trial, (171) clinicians in hospitals randomised to the intervention were encouraged to use the OARs as part of their clinical assessment.

Table 3.5. Results of studies of serious bacterial infection in children with fever

Outcomes	Number of trials/N of trial or comparison	Result	
		Intervention	Control
Process outcomes			
Length of stay in the ED (median minutes (25 th to 75 th percentile))	2 (Roukema et al., 2008 (198))*/164 (de Vos-Kerkhof et al., 2015 (204))/439	138 (104 to 181) 117 (84 to 158)	123 (83 to 179)(p=0.16)† 114 (81 to 162)(p>0.05)†
Clinician decisions/appropriateness of decisions			
Antibiotic prescriptions	1 (Lacroix et al., 2014 (205))*/271	Clinician + CPR 41% CPR alone 31%	42% (p=0.88) 42% (p=0.009) in entire study cohort
Laboratory tests ordered when recommended by CPR (%)	1 (Roukema et al., 2008 (198))/164	82%‡	44% (p<0.002)†
Chest x-ray performed when no pneumonia (false positive)	1 (de Vos-Kerkhof et al., 2015 (204))*/439	60%	57% (p>0.05)
No chest x-ray performed when pneumonia (false negative)	1 (de Vos-Kerkhof et al., 2015 (204))*/439	1%	1% (p>0.05)
Urine culture performed when no UTI (false positive)	1 (de Vos-Kerkhof et al., 2015 (204))*/439	67%	53% (p>0.05)
No urine culture performed when UTI (false negative)	1 (de Vos-Kerkhof et al., 2015 (204))*/439	0%	0.5% (p>0.05)
Urine dipstick performed	1 (de Vos-Kerkhof et al., 2015 (204))/439	71%	61% (p>0.05)
Overall diagnostics (minus urine dipstick analysis)	1 (de Vos-Kerkhof et al., 2015 (204))/439	57%	63% (p>0.05)
Antibiotics prescribed at discharge	1 (de Vos-Kerkhof et al., 2015 (204))/439	32%	36% (p>0.05)
% with SBI prescribed antibiotics	1 (de Vos-Kerkhof et al., 2015 (204))/439	93%	93% (p>0.05)
% with no SBI prescribed antibiotics	1 (de Vos-Kerkhof et al., 2015 (204))/439	23%	27% (p>0.05)
Hospitalisation	2 (de Vos-Kerkhof et al., 2015 (204))/439 (Lacroix et al., 2014 (205))/271	12% 34%	11% (p>0.05) 36% (p=0.81)
Accuracy			
Area under the curve for any SBI	1 (Lacroix et al., 2014 (205))/271	0.91 (95%CI 0.87 to 0.95)	
Area under the curve for pneumonia	1 (de Vos-Kerkhof et al., 2015 (204))/439	0.83 (95%CI 0.75 to 0.90)	
Area under the curve for other SBI	1 (de Vos-Kerkhof et al., 2015 (204))/439	0.81 (95%CI 0.72 to 0.90)	

CPR – clinical prediction rule; SBI – serious bacterial infection; UTI – urinary tract infection

*primary outcome; †p value for difference between intervention and control groups; ‡adjusted for age

When used as a triage test, the OARs did not decrease total length of stay in the emergency department (mean difference -6.7 minutes 95%CI -20.9 to 7.4), and there was no difference in patient satisfaction ratings and radiography requests between the study groups. (188) The OARs, when used and applied at the discretion of the clinician as an add-on test, significantly decreased radiography requests (76% vs 99% p=0.03). Three fractures were later diagnosed among participants who had not received an x-ray - all were randomised to the experimental arm. However, there was no active follow-up in this trial and participants were only advised to

consult again if there was persistent pain or inability to walk so it is possible other fractures may have been missed. In this trial, 96% and 98% of patients in the experimental and control groups respectively were satisfied with the care received. (171)

Table 3.6. Results of studies of ankle or mid-foot fracture by outcome

Outcomes	Number of trials/N of trial or comparison	Result	
		Intervention	Control
Patient outcomes			
Patients satisfied with care (%)	1 (Auleley et al., 1997 (171))/1911	96%	98%
Patient satisfaction rating (median score on 5 point scale)	1 (Fan et al., 2006 (188))/124	4	4 p=0.34 [†]
Process outcomes			
Total length of stay in ED (minutes)	1 (Fan et al., 2006 (188)*)/124	MD -6.7 (95%CI -20.9 to 7.4)	
Time from triage to registration (mean minutes)		8.0	8.0 p=0.80 [†]
Time from registration to room assignment (mean minutes)		20.0	13.0 p=0.05 [†]
Time from room assignment to clinician assessment (mean minutes)		25.0	19.0 p=0.16 [†]
Time from clinician assessment to disposition (mean minutes)		21.0	27.0 p=0.62
Clinician decisions/appropriateness of decisions			
Radiography requests (%)	2 (Auleley et al., 1997 (171)*)/1911	76%	99% p=0.03 [†]
	(Fan et al., 2006 (188))/124	94%	89% p=0.36 [†]
Number of fractures in those not x-rayed	1 (Auleley et al., 1997 (171))/1911	3	0 [†]
Use and implementation/application outcomes [§]			
Use of the data collection for containing the CPR	1 (Auleley et al.,1997 (171))/1911	75%	
Requested x-ray when CPR recommended x-ray		99%	
Requested x-ray when CPR recommended no x-ray		21%	

CPR – clinical prediction rule; MD – mean difference; ED – emergency department

*primary outcome; [†]p value for difference between intervention and control groups; [‡]There was no active follow-up of participants in this trial and individuals with persistent pain or difficulty walking may have presented for care elsewhere; [§]only relevant to the intervention arm of studies where use of the CPR was discretionary

Studies of acute coronary syndromes (Table 3.7)

Two studies judged to be at low risk of bias evaluated different diagnostic strategies incorporating different CPRs and tests for individuals presenting with possible cardiac chest pain. (60, 201) In the first of these studies, a strategy incorporating a diagnostic CPR (TIMI score), ECG and cardiac troponin testing significantly increased the number of patients successfully discharged within 6 hours (discharge was considered successful if it occurred within 6 hours of emergency department arrival and patient did not experience a major cardiac adverse event within 30 days) (OR 1.92 95% CI 1.18 to 3.13), (201) compared with a conventional chest pain protocol. In the second study, a diagnostic strategy comprising a CPR and NT-proBNP test decreased the number of patients hospitalised (OR 0.6 95% CI 0.4 to 0.9) with no differences in death or myocardial infarction between study groups after 1 year follow-up. (60)

Table 3.7. Results of studies of acute coronary syndrome by outcome

Outcomes	Number of trials/N of trial or comparison	Result	
		Intervention	Control
Patient outcomes			
Death or myocardial infarction at 1 year follow-up	1 (Sanchis et al., 2009 (60))/320	HR 1.9 (95%CI 0.7 to 5.2)	
Major adverse cardiac event at 30 day follow-up	1 (Than et al., 2014(201))/542	1	0
Process outcomes			
Revascularisations at index visit (%)	1 (Sanchis et al., 2009 (60))/320	8.1%	18.1% p=0.01 [†]
Urgent post discharge revascularisations (%)		1.3%	2.5% p=0.07 [†]
Planned post discharge revascularisations (%)		5%	0.6% p=0.04 [†]
Clinicians' decisions/appropriateness of clinicians' decisions			
Discharge within 6 hours without major adverse cardiac event within 30 days	1 (Than et al., 2014 (201)*)/542	OR 1.92 (95%CI 1.2 to 3.1)	
Hospitalisation at index episode	1 (Sanchis et al., 2009 (60)*)/320	OR 0.6 (95%CI 0.4 to 0.9)	
Use and implementation/application outcomes [‡]			
Number (%) classified as low risk by the diagnostic pathway incorporating CPR but admitted to hospital	1 (Than et al., 2014 (201))/542	35 (12.9%)§	

HR – hazard ratio; OR – odds ratio; CPR – clinical prediction rule

*primary outcome; †p value for difference between intervention and control groups; ‡only relevant to the intervention arm of studies where use of the CPR was discretionary; §none received a diagnosis of acute coronary syndrome

Studies of single conditions (Table 3.8)*Clinical outcomes*

Two studies reported clinical outcomes as the primary outcome of the study. In the first, a score directed treatment algorithm for patients with upper abdominal complaints significantly decreased symptom severity (MD on a scale of 0-10 2.5 95% CI 1.49 to 3.51). (190) The other study was an equivalence trial of a triage strategy of 'bedside tests' incorporating a diagnostic CPR for directing ventilation perfusion scanning verses a strategy of scanning all patients. In this study there was no significant difference in venous thrombotic events among patients not taking anticoagulation agents during follow-up (% difference in venous thromboembolic event rate -0.6 95% CI -4.1 to 2.9), but the triage strategy excluded pulmonary embolism in 34% of patients (who therefore avoided ventilation perfusion scanning) and reduced other diagnostic imaging test performed. (197)

Clinicians' decisions/appropriateness of clinicians' decisions

Prescriptions for antibiotics were significantly reduced with use of a CPR in patients with non-severe community acquired pneumonia of unknown aetiology, and in patients with suspected pneumonia, (194) with no difference in unfavourable outcomes between interventions. (202) Radiography requests increased, but time in the emergency department significantly decreased, in a study of a CPR used as a triage test in children with extremity trauma. (191) There was no difference in appropriate referrals in adults with pigmented skin lesions with use of a diagnostic protocol incorporating a CPR and the MoleMate scanning technique. (203) A CPR for head injured patients did not lead to a reduction in ED use of CT imaging, (200) but a CPR for patients with blunt head and neck trauma and possible cervical spine fracture significantly decreased imaging without missing injuries. (199)

Accuracy

A CPR had similar sensitivity and higher specificity for complete bowel obstruction in patients with acute small bowel obstruction compared to contrast radiography. (186)

Table 3.8. Results of single studies of different clinical conditions

Trial/N of trial	Outcomes	Result	
		Intervention	Control
Children with pneumonia of bacterial aetiology			
Torres et al., 2014 (202)/120	Antibiotic prescriptions* Unfavourable clinical outcome		OR 0.13 (95%CI 0.05 to 0.35) OR 1.0 (95%CI 0.2 to 3.6)
Pneumonia			
McGinn et al., 2013 (194)/395	Antibiotic prescriptions* Chest radiographs ordered Decision support tool opened CPR calculator opened		RR 0.79 (95%CI 0.64 to 0.98) [†] RR 0.98 (95%CI 0.60 to 1.62) [†] 42.5% 41.5%
Children with joint or bone injury of the extremities			
Klassen et al., 1993 (191)/991‡	Radiography requests* CPR only CPR as triage test Missed fractures (%) CPR only CPR as triage test Time spent in ED (mean hrs)		RR 0.94 (95%CI 0.89 to 0.99) RR 1.12 (95%CI 1.08 to 1.16) 3.2% 0% 3.3 (SD 1.7)
Suspicious pigmented lesions			
Walter et al., 2012 (203)/1580 [‡]	Appropriate referral rate [¶] Benign lesions appropriately managed in PC (%) Sensitivity Specificity Lesions referred (%)		% difference -8.1 (95%CI -18.0 to 1.8) % difference 0.5 (95%CI -0.6 to 2.0) 98.5% 84.4% 29.8%
Pulmonary embolism			
Rodger et al., 2006 (197)/398	Venous thromboembolic event rate during 3 month f-up* Total bleeding episodes during 3 months follow-up Diagnostic imaging tests performed (mean no./patient)		% difference -0.6 (95%CI -4.1 to 2.9) 2 (p=0.45)§ 1.9 (p<0.001)§
Gastro-oesophageal reflux disease			
Horowitz et al., 2007 (190)/132	Relief of symptoms* Improvement in daily activities Management costs (US\$) Number of GP visits Number of specialist referrals Cost of medications (US\$) Number of imaging diagnostic tests		MD 2.5 (95%CI 1.49 to 3.51) MD 0.6 (95%CI 0.18 to 1.02) MD -138.00 (95%CI -230.70 to -45.30) MD -0.3 (95%CI -0.59 to -0.00) MD -0.36 (95%CI -0.65 to -0.09) MD -68.00 (95%CI -269.40 to 133.38) MD -0.10 (95%CI -0.23 to 0.03)
Complete acute bowel obstruction			
Bogusevicius et al., 2002 (186)/80	Sensitivity* Specificity* Time to make diagnosis (hrs) Mortality	100% (95%CI 86.2 to 1) 87.5% (95%CI 6.39 to 96.5) 1 5.0%	100% (95%CI 87.5 to 1) 76.9% (95%CI 49.7 to 91.8) 16 (p<0.001)§ 0.0%
Clinically important brain injury			
Stiell et al., 2010 (200)/4531	% change in CT scan rates from before to after* Missed brain injuries during the ED visit % of clinicians using rule % of clinicians accurately completing rule	13.3% (95%CI 9.7 to 17.0) 0 78% 82.5%	6.7% (95%CI 2.6 to 10.8) (p=0.16)** 0
Cervical spine fracture			
Stiell et al., 2009 (199)/11,824	% change in cervical spine imaging before to after* Missed cervical spine fractures during the ED visit % of clinicians using rule % of clinicians accurately completing rule	-12.8 (95%CI -9.2 to -16.3) 0 85.7% 82.9%	12.5 (95%CI 7.2 to 18.2)(p=0.00)** 0

CPR – clinical prediction rule; SD – standard deviation; MD – mean difference; OR – odds ratio; RR – risk ratio

*primary outcome; [†]adjusted for age; [‡] 991 injury sites in 974 participants; [§]p value for difference between intervention and control group; ^{||} 1580 lesions in 1297 participants; [¶]no. of referred lesions secondary care experts decided to biopsy or monitor/number referred; ^{**}p value for relative change in mean imaging rates from the before period to the after period between intervention and control hospitals.

Effect on patient outcomes across clinical conditions

Three studies reported patient outcomes as the primary outcome of the study. (182, 190, 197) Two of these studies reported decreased patient symptoms of sore throat or upper gastrointestinal discomfort with application of a diagnostic strategy incorporating a CPR compared to a strategy of delayed antibiotic prescribing or usual clinical management respectively. (182, 190) The third study compared a triage strategy incorporating a diagnostic CPR with other bedside tests and selective VQ scan, with a strategy of ventilation perfusion scanning in all patients with suspected pulmonary embolism. (197) This equivalence study reported similar rates of venous thromboembolic events (primary outcome) in both study groups but reduced use of diagnostic tests.

3.5.5 Assessment of the reporting of interventions

Assessment of the reporting of interventions in these studies found details of the components necessary to replicate the intervention were often missing (Table 3.9). When present, they were usually only partly described (for example, a diagnostic strategy comprising a CPR and laboratory test described the technique used to perform the laboratory test but did not describe the CPR), or were described with very low level of detail. Only 1 of 25 included studies was judged to have described the diagnostic strategy and criteria for arriving at a diagnosis or decision in both the experimental and control groups, (60) and 6 studies to have reported both whether training or exposure to the CPR was provided and the means by which the CPR was implemented into the workflow. (171, 188, 194, 198, 203, 204) Control interventions that were variably described as 'usual' care were less frequently described than control interventions comprising more technological procedures (e.g. contrast radiography). Most studies did not report providing clinicians with briefing or training in use of the CPR, or details on the manner in which it was integrated into workflow. In 6 studies, the CPR was a computer based tool. In one of these it was integrated into the electronic health record platform, (184, 186, 194, 198, 203, 204) and in 2 studies it was a paper based tool. (195, 196) In the remainder, the means by which the CPR was made available could not be determined.

Table 3.9. Minimum required elements for reporting of diagnostic strategies and implementation methods

Study	Primary outcome	Study arm	Intervention description		Implementation description		
			Diagnostic tests being evaluated are described?	Criteria for arriving at a diagnosis, testing or treatment decision are described?	Selection of treatment and how it is administered is described?	Training of clinicians in use of the CPR or application of its output is provided, is and if provided, is described?	Method by which the CPR is made available for use is described?
Group A streptococcus throat infection							
Worrall et al., 2007 (183)	Antibiotic prescribing	Experimental Control	Yes No	Yes No	NA NA	No No	No Yes
McIsaac & Goel 1998 (195)	Antibiotic prescribing	Experimental Control	Yes Yes	No Yes	NA NA	No No	Yes Yes
McIsaac et al., 2002 (196)	Unnecessary antibiotic prescribing	Experimental Control	Yes Yes	Yes No	NA NA	No Yes	Yes Yes
McGinn et al., 2013 (194)	Antibiotic prescribing	Experimental Control	Yes No	Yes No	NA NA	Yes No	Yes No
Little et al., 2013 (182)	Symptom severity	Experimental Control	Yes Yes	Yes No	No No	No No	No No
Acute appendicitis							
Douglas et al., 2000 (187)	Time to therapeutic operation	Experimental Control	Yes No	Yes No	NA NA	No No	No No
Farahnak et al., 2007 (189)	Time to surgery	Experimental Control	Yes No	Yes No	NA NA	No Yes	No No
Lintula et al., 2010 (193)	Diagnostic accuracy	Experimental Control	Yes Yes	No Yes	NA NA	Yes Yes	No No
Lintula et al., 2009 (192)	Diagnostic accuracy	Experimental Control	Yes Yes	Yes No	NA NA	Yes No	No Yes
Wellwood et al., 1992 (184)	Diagnostic accuracy	Experimental Control	No No	No No	NA NA	No No	Yes Yes
Serious bacterial infection in children with fever							
Roukema et al., 2008 (198)	Length of stay	Experimental Control	Yes No	Yes No	NA NA	Yes Yes	Yes No
Lacroix et al., 2014 (205)	Antibiotic prescribing	Experimental Control	Yes Yes	Yes No	NA NA	Yes Yes	No Yes
de Vos-Kerkhof et al., 2015 (204)	Appropriate test use	Experimental Control	Yes No	Yes No	NA No	Yes No	Yes Yes

Table 3.9. Continued

Study	Primary outcome	Study arm	Intervention description			Implementation description		
			Diagnostic tests being evaluated are described?	Criteria for arriving at a diagnosis, testing or treatment decision are described?	Selection of treatment and how it is administered is described?	Training of clinicians in use of the CPR or application of its output is provided and if provided, is described?	Method by which the CPR is made available for use is described?	
Ankle or mid-foot fracture								
Auleley et al., 1997 (171)	Radiography requests	Experimental Control	No No	No No	NA NA	Yes Yes	Yes Yes	
Fan et al., 2006 (188)	Length of stay in the emergency department	Experimental Control	No No	No No	NA NA	Yes Yes	Yes Yes	
Acute coronary syndromes								
Than et al., 2014 (201)	Safe discharge	Experimental Control	Yes Yes	Yes No	NA NA	No No	No No	
Sanchis et al., 2010 (60)	Hospitalisation	Experimental Control	Yes Yes	Yes Yes	NA NA	No No	No No	
Single studies of different clinical conditions								
Torres et al., 2014 (202)	Antibiotic prescribing	Experimental Control	Yes Yes	Yes No	NA NA	No Yes	No Yes	
McGinn et al., 2013 (194)	Antibiotic prescribing	Experimental Control	Yes No	Yes No	NA NA	Yes Yes	Yes No	
Klassen et al., 1993 (191)	Radiography requests	Experimental Control	Yes No	Yes No	NA NA	Yes Yes	No Yes	
Walter et al., 2012 (203)	Appropriate referral	Experimental Control	Yes Yes	Yes No	NA NA	Yes No	Yes No	
Rodger et al., 2006 (197)	Venous thromboembolic events	Experimental Control	Yes No	Yes No	NA NA	Yes No	Yes No	
Horowitz et al., 2007 (190)	Symptom relief	Experimental Control	Yes No	Yes No	NA NA	No No	No No	
Bogusevicius et al., 2002 (186)	Diagnostic accuracy	Experimental Control	Yes Yes	No Yes	NA NA	No Yes	Yes No	
Stiell et al., 2010 (200)	Computed tomography scan rates	Experimental Control	Yes No	Yes No	NA NA	Yes No	No No	
Stiell et al., 2009 (199)	Cervical spine imaging rates	Experimental Control	Yes No	Yes No	NA NA	Yes No	No No	

3.6 Discussion

Diagnostic CPRs evaluated in this review were found to have beneficial effects on process outcomes in some clinical conditions, and in some cases had a positive effect on patient health. Though improvement in patient outcome, or increased efficiency of the diagnostic process without worsening patient outcomes, is the ultimate measure of effectiveness for diagnostic CPRs, few included studies primarily aimed to determine the effect of diagnostic CPRs, or diagnostic strategies incorporating CPRs, on patient outcomes. The majority of studies included in this review investigated the consequences of use of the CPR on clinicians' decisions to test or treat. Study methods, intervention and implementation details necessary for interpretation and safe application of the intervention were generally poorly reported. This non-transparency also hinders attempts to replicate studies or their findings, and erodes the value of the research in this area. (206, 207)

The conclusions drawn in this review are based on a small number of studies and, as a substantial number of these studies were categorised as high or unclear risk of bias on 3 or more domains of bias (one of which was performance bias), caution is advised in interpretation of their results. Potential bias varied across studies for different clinical conditions and it was often difficult to judge whether any bias would result in an over or under estimation of treatment effect. The assessment of risk of bias relied upon the reporting of trials and there was insufficient detail to confidently assess risk in many cases. Due to the nature of the experimental intervention, which requires interaction with and interpretation by clinicians, blinding of clinicians was not possible. As such, clinicians' prior expectations of effectiveness of the intervention were judged to have the potential to lead to bias either through disparity in other care that is administered to patients or by affecting clinicians' decisions which are an outcome in many studies. The risk arising from non-blinding of individuals assessing outcomes was judged unclear for most studies. This was due either to inadequate reporting, or absence of blinding of independent data collectors or adjudicators.

The comparative performance of the interventions evaluated in this review, and the means by which performance is judged, are likely to be dependent on the context in which the diagnostic strategy is implemented. For example, in a study of the effect of a CPR for predicting streptococcal infection conducted in general practice in the UK, discretionary application of the CPR decreased the severity of sore throat symptoms compared to a strategy of delayed antibiotic prescribing. (182) It is not clear how the prediction rule could affect resolution of sore throat, but it is possible that it helped to identify patients who would respond to antibiotics more accurately or quickly. In a setting where antibiotics are prescribed for sore

throat more frequently and possibly earlier, the relative effect of the prediction rule on symptoms may be different. As another example, a diagnostic strategy incorporating a CPR and BNP testing may lead to less hospital admissions in a country where admissions for possible cardiac chest pain are common, (60) but may have less effect in a situation where admissions are infrequent.

Conclusions of many of the studies included in this review are likely to be limited by inadequate sample size. Test-treatment trials reporting patient outcomes may need to be considerably larger than standard treatment trials in order to account for the fact that the effects of administered treatments are only experienced in the small proportion of patients who receive different care as a result of their diagnosis. (208) However, this need for inflation of the sample size may depend on the mechanism by which the CPR operates. If the effects on health outcomes are the result of changes to the timing of tests and treatments these are likely to be experienced by all randomised participants rather than only the small proportion of patients managed differently. Another reason why a larger sample size may be required is the high rate of contamination of aspects of the CPR intervention in the control group when randomisation is at the patient level. Trial designs that attempt to mitigate this by randomising groups of people rather than individuals will also need inflation of required sample sizes to retain equivalent power to an individually randomised trial after allowance is made for correlation of observations within the patient clusters. (209) Of the included studies utilising this design, only half took into account the clustered design in sample size calculations.

Within and across clinical conditions, there was heterogeneity in the degree to which CPRs and diagnostic strategies incorporating CPRs were used and their output applied. The protocols of the included studies took one of two approaches to the use of the CPRs or diagnostic strategies: a pragmatic approach in which clinicians could decide whether or not to use the tool, or an approach in which clinicians in the study were expected to use the CPR or were provided with the output from it. Further, there were varying degrees to which the clinician was required to follow the recommendation provided by the directive rules or strategies. In some studies, the subsequent treatment provided was dictated entirely by the CPR. In others, the clinician was 'encouraged' to follow the recommendations, and in others, clinicians could adopt or ignore the recommendations at their discretion. These variations may lead to differences in intervention effect but also have implications for transferring the research findings to clinical practice. Results from studies mandating use of a CPR and carefully monitoring its correct application may be different to results seen when the CPR is introduced

in a situation where clinicians are given license to override its recommendations. It has been suggested that impact studies should assess both *actual* impact – impact when clinicians can use their discretion in following the CPRs recommendations, and *potential* impact – measured by analysing the CPRs recommended decision regardless of implementation. (5) This was done by 2 studies included in the review. In one of these studies, (191) application of the CPR alone reduced x-rays but missed more fractures than standard care. When the CPR was used as a triage test and those negative on the CPR were assessed by clinicians, far more x-rays were performed, but no fractures were missed. In the other study, (205) strictly following the CPR recommendations would have resulted in a treatment rate (antibiotic prescribing) of 31% as opposed to an actual treatment rate of 42% when clinicians could override the CPR recommendations (control group treatment rate 42.1% and 41.7% in the entire study cohort). Such information is likely to assist in the interpretation of impact study findings by informing of the interactions taking place between the clinician and the CPR and the reasons for any disagreements.

The design of another trial invites discussion on whether CPRs may affect clinicians' judgment in a particular clinical situation through refinement of the process of data collection alone or through the combination of refined data collection and probability information. In this 3-arm study, the effects of clinicians' use of an assistive CPR were compared to the effects of a) clinicians' assessment with a standardised data collection form and, b) clinicians' usual care with no diagnostic aid. (184) For the outcomes studied, clinicians with structured data collection forms and clinicians with structured data collection and a probability estimate, generally performed as well as each other and better than clinicians with no diagnostic aids. The issue of whether, and in what circumstances improved performance occurs through the process of refining clinicians' judgment or the combination of refinement and information in the form of probability estimates and or recommendations, deserves further research.

Though clear reporting of interventions is necessary for interpretation of study findings and safe replication of the intervention in practice, (210) documentation of the interventions tested in the majority of included studies was poor. This is similar to research on other complex interventions and non-pharmaceutical treatments. (211, 212) Furthermore, studies rarely stated how a diagnostic CPR is expected to alter outcomes relative to the alternate diagnostic strategy, making it difficult to judge the adequacy of the outcomes reported. For many of the included trials where the control intervention could broadly be described as 'usual' care, no, or minimal description of the test method or the criteria by which

management decisions were made was provided. 'Usual' care had various permutations ranging from what is termed 'wild type' or 'care as it is now', to a more regimented guideline driven care. (213, 214) It is acknowledged that such strategies are internalised, likely complex, probably highly variable and nuanced and difficult to translate into a prescriptive format. However, lack of even mention of the tests performed makes it very difficult to interpret differences in trial outcomes and to judge generalisability. Furthermore, basic details about the process of implementation were infrequently reported making it difficult to know whether failure to demonstrate an effect is more likely due to inadequate implementation of the experimental strategy, than lack of effect of the experimental intervention itself. Nor did the majority of trials provide information on which to judge the risk of behaviour change among clinicians in the control groups arising from knowledge of study conditions. (214)

The information provided by a diagnostic prediction rule should either: 1) lead ultimately to improved outcomes for patients (e.g. resolution of symptoms), or 2) maintain (not worsen) patient outcomes whilst providing other benefits (e.g. reduced unnecessary tests or treatments). Despite this, few studies included in this review report any patient outcomes, and fewer still report patient outcomes as the primary study outcome. Clinicians' decisions to test or treat are the primary outcomes, and sometimes the only study outcome, in the majority of included studies. There may be several explanations for this: randomised studies reporting patient outcomes are difficult, costly and time consuming to conduct; researchers may believe that patient management is a valid surrogate for health outcomes; or researchers may select outcomes that reflect the primary intention of many diagnostic CPRs to reduce testing or treatment. However, recent research suggests that it is not possible to infer the effects of a diagnostic test on patient outcomes based on how a test influences management decisions. In an analysis of a large sample of diagnostic randomised controlled trials, the effects of the index test on further diagnostic and therapeutic interventions did not correlate with the effects on patient outcomes. (172) This study also found that estimates of accuracy do not inform well about the clinical utility of diagnostic tests. Given the multitude of ways CPRs may affect patient outcomes, (50) improved accuracy or management decisions afforded by a CPR are neither a necessary requisite nor a guarantee for improving patient health. Though measurement of the effects of CPRs on patient management may be of some use for planning further evaluations of a CPR, and as part of a suite of outcome measures to assist in understanding the means by which a CPR may exert its effects, we argue that impact studies reporting only management decisions, or reporting management decisions without

considering effects on patient outcomes, are insufficient to judge the clinical utility of diagnostic CPRs.

To our knowledge, this is the first review of diagnostic CPRs across a range of clinical conditions. We are aware of one systematic review evaluating the impact of diagnostic prediction rules for acute appendicitis. (170) This review included only 1 trial, (184) which we also included in our review (along with 4 other trials subsequently published). Based on the findings of this trial which found similar and statistically significantly decreased admission rates among clinicians randomised to a) an assistive CPR and b) standardised data collection forms, compared to clinicians with no diagnostic aid, the review concluded that standardised data collection is a promising strategy for assisting diagnosis, while criticising the reporting and conduct of existing studies. The findings of our review, that CPRs reduce prescribing and test ordering for some conditions, are broadly consistent with existing research evaluating clinical decision support tools which has found that some systems can improve test ordering and antibiotic prescribing behaviour. (173-175)

Our review has limitations. Because of the large number of titles and abstracts retrieved in the searches, only 1 reviewer performed screening of titles and abstracts, with a second reviewer screening only a proportion. Therefore, some studies may have been overlooked. However, screening of systematic reviews of clinical decision support systems, reference checks and forward searches minimised the possibility that eligible studies were missed. The presence of study publication bias in this review is possible. For instance, many of the CPRs were tested by the researchers who developed the CPR and thus may be more likely to submit studies with positive results for publication. In reporting whether a study described components of the interventions, we determined only whether a description was present, rather than providing a judgment about the adequacy of the description. Consequently the review is likely to overestimate the reporting quality of the included studies, as components were judged to have been described even if only partially so, or with little detail. Furthermore, the criteria assessed were considered by the authors to be the minimum essential to the reporting of intervention content. To properly appraise reporting quality of impact studies, more criteria should be considered.

3.7 Conclusion

This review provides insight into the current status of research evaluating the impact of diagnostic clinical prediction rules and provides information that may assist clinicians and policy makers' decisions regarding the application of these tools. This review found that

diagnostic CPRs improve process of care measures for some clinical conditions, and in some cases improved or maintained patient health while providing other benefits. However, this conclusion is based on a small number of studies, many of which are judged to be at high or unclear risk of bias and is likely to be context dependent. It is apparent from this review that future impact studies need to be more carefully designed and conducted and more thoroughly reported. Consideration of the many mechanisms by which a CPR may alter outcomes during the trial design stage, should guide the nature and number of outcomes measured and facilitate understanding of why particular effects are observed. Use of a framework such as that developed by Ferrante de Ruffano and colleagues (50) may assist firstly in identifying the means by which a CPR may alter the existing diagnostic pathway, and secondly to consideration of the full range of direct and downstream outcomes that should be measured. Furthermore, reporting of such studies should be improved to assist interpretation and replication in practice. Establishing benefit to patient health or showing that patient health can be maintained while providing other benefits, should be the priority of impact evaluations of diagnostic prediction rules.

3.7.1 Acknowledgements

We would like to thank Elaine Beller for her assistance with the meta-analyses.

3.7.2 Author contributions

Conceived the experiment: SS. Designed the experiment: SS, JD, PG. Analysed the data: SS, JR, KB. Wrote the first draft of the manuscript: SS. Contributed to writing of the manuscript: SS, JD, KB, JR, PG. Agree with manuscript results and conclusions: SS, JD, KB, JR, PG.

SS, JD and KB positions receive funding from the Screening and diagnostic Test Evaluation Program which is supported by a National Health and Medical Research Council Program Grant (<https://www.nhmrc.gov.au>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 4 The identification of serious bacterial infection in children with fever presenting to primary care

4.1 Preface to Chapter 4

The following chapters of this thesis focus on diagnostic prediction rules in a specific clinical area where diagnosis is known to be challenging, the identification of serious bacterial infection in children. The goal of this part of the thesis is to assist primary care clinicians' diagnostic management of children with possible serious infection through the derivation and validation of a diagnostic prediction rule, and simplified versions of the rule, and elucidation of the accuracy, independent and added value of the inflammatory biomarker C-reactive protein. The clinical problem, the rationale for, and design of proposed studies to achieve this goal are outlined in this chapter. Reasons for inability to complete two of the proposed projects, and subsequent modifications to one of the proposed studies are also discussed.

4.2 The clinical problem

The investigation of fever in young children has been identified by physicians in several countries as one of the most difficult clinical problems. (136) Fever is also one of the most common reasons for a child to be seen by primary care practitioners (215) and to present to emergency departments for care. (216-218) However, differentiating a child with serious infection from the large number of children who have minor illnesses that are often self-limiting is difficult for several reasons. Firstly, serious infections are rare in most settings. Incidences of <1% have been reported in primary care (219) and 7% in a large cohort of febrile children presenting to an Australian emergency department. (220) Secondly, infection is a dynamic process and children may present in the early stages of disease when severity is not apparent and may then deteriorate rapidly. Thirdly, the initial presentations of serious and non-serious infections can be similar. (221) Fourthly, in some settings, potentially useful diagnostic procedures may not be available. Lastly, assessment of children can be difficult and is often undertaken by staff with limited paediatric experience or under high pressure because of large patient volumes. (222)

Clinical assessment, including history taking for symptoms and physical examination for signs, forms the basis for diagnostic decision making in children with possible serious infection. However, the consideration of signs and symptoms alone often results in residual diagnostic uncertainty. (223) While some clinical features increase or decrease the probability that a child has serious infection, none is sufficient on its own to substantially raise or lower the risk of serious infection. (223) Though some features are recognised as red flags or highly specific warning signs for serious infection, (223) these features are seen infrequently in the primary care setting (224) and are uncommon even in children with serious infection. (222)

Due to the small but real chance that the child with acute infection will not recover as expected and will suffer complications with severe and occasionally fatal consequences, (225) many children who ultimately have self-limiting diseases are referred to secondary care as a precaution. (226) This use of secondary care resources results in significant cost to the healthcare system and subjects the child to risks inherent in further diagnostic testing and or treatment. Prescribing antibiotics in the hope of preventing a negative outcome is another way primary care clinicians may deal with this uncertainty. (227, 228) This low threshold for prescribing exposes many children to unnecessary adverse effects and costs and contributes to antibiotic resistance at an individual and societal level. (229, 230) On the other hand, children who do have a serious bacterial infection may not be sent to hospital promptly when the first

signs of severe infection are present, or are prescribed antibiotic treatment when they would likely benefit from them. (220, 221)

4.3 Strategies to improve the diagnostic management of children with possible serious bacterial infection presenting to primary care

In an effort to bridge the diagnostic gap between the predictive value achievable by the consideration of clinical features and the implicit threshold for referral or treatment, clinicians may rely on their instinct that something is wrong, utilise diagnostic safety netting procedures, (226) or conduct more invasive and involved testing. Clinical prediction models that combine signs and symptoms and laboratory or imaging tests have also been developed as a means to assist clinicians to identify children with serious bacterial infections. Two of these strategies, diagnostic prediction rules and additional diagnostic testing will be considered further in this chapter.

4.3.1 Clinical prediction rules as a strategy to improve the diagnostic management of children with possible serious bacterial infection

Many clinical prediction rules have been proposed to differentiate children with a serious bacterial infection from those with a benign self-limiting non-bacterial infection or to identify specific serious infections such as pneumonia, meningitis or septicaemia. (222) The expectation, though rarely stated in studies deriving these rules, is that clinicians' use of the prediction rule and/or knowledge of the estimated probability of serious bacterial infection, will improve the accuracy of the clinicians' diagnosis leading to improved management decisions. Ideally, the prediction rule will improve the sensitivity of the clinicians' diagnosis, thus avoiding missed cases of serious bacterial infection, as well as the specificity of diagnosis, thus minimising the number of children without serious bacterial infection subject to further testing or treatment. Though the primary mechanism by which diagnostic prediction rules may be expected to drive changes to decisions and downstream patient health is through improved accuracy, prediction rules may also alter the temporality of clinicians' decisions and actions. Use of the prediction rule, which focuses the clinician on a small number of highly diagnostic indicators for serious bacterial infection, may lead to earlier diagnosis and subsequently earlier management decisions.

4.3.2 Additional testing

In some clinical settings clinicians may seek to increase diagnostic certainty after physical examination and history taking through additional diagnostic testing. These follow-up tests are often more invasive, expensive, inconvenient and carry a greater risk of harm. Tests that may be utilised during the diagnostic workup of a child with possible serious infection include blood

tests for white blood cell counts and inflammatory markers, urinalysis, lumbar puncture as well as imaging. (231) Inflammatory markers are proteins whose concentrations in the blood increases or decreases during inflammation and are seen to play an important role in identifying the presence, absence or severity of the inflammatory response in individuals suspected of having serious infection. One such inflammatory biomarker is C-reactive protein.

4.3.2.1 The inflammatory biomarker C-reactive protein

C-reactive protein (CRP) has been utilised as a test for differentiating between bacterial and viral infection because it has been shown that a very high CRP ($>100\text{mg/L}$) is more likely to occur in bacterial than viral infections, and a normal CRP is unlikely in the presence of significant bacterial infection. However, intermediate CRP concentrations may be found in both bacterial and viral conditions and low concentrations ($>20\text{mg/L}$) have been associated with increased risk of pneumonia. (232, 233)

Although CRP testing is primarily carried out in a laboratory in a hospital setting, the availability of point of care tests which provide immediate results, and use finger stick droplets rather than large aliquots of blood, suggests that CRP testing may have a role in primary care settings. In the early stages of this thesis, CRP was widely used as a point of care test in primary care in some European countries, and its use was increasing, or being promoted in other developed countries. (234, 235)

4.4 Proposed studies intended to assist clinicians' diagnostic management of children with possible serious bacterial infection presenting to primary care

4.4.1 The derivation and validation of a clinical prediction rule

In the early stages of this thesis (to end 2007), several prediction rules based on clinical features from the patients' history or physical examination had been developed for estimating the risk of serious bacterial infections overall (236, 237) and specific serious bacterial infections. (238-243) The prediction rules available at the time are described in Table 4.1. They were, however, of questionable appropriateness for clinical practice. Some had key methodological limitations (e.g. predictors were not identified and combined using multivariable statistical techniques), (237, 240, 243) or were limited to use in very young age groups (e.g. babies). (238, 242) Others had been derived in tertiary settings or in populations with a higher incidence of serious bacterial infection prior to the introduction of routine vaccination against *Haemophilus influenzae* type b and *Streptococcus pneumoniae*. (236, 237) Further to this, only one of these models had been externally validated (237) and in the

validation studies, the potential of the tool to rule out serious infection was variable when used in conjunction with laboratory tests. (244-246)

Table 4.1. Clinical prediction rules comprised of predictors from history or physical examination for serious bacterial infection overall or specific serious bacterial infections in children (published to end 2007)

Clinical prediction rule	Predictors included in the rule	Methodological and/or generalisability limitations
All serious infections		
Bleeker et al., 2001 (236)	Duration of fever, history of poor micturition, history of vomiting, age >1 year, temperature, chest wall retractions, poor peripheral circulation	Derived in children <36 months Derived in tertiary paediatric hospital with prevalence of serious bacterial infection 25%
Yale observation scale McCarthy et al., 1982 (237)	Quality of cry, reaction to parents', state variation, colour, state of hydration, response to social overtures	Derived in children <24 months Derived in tertiary paediatric hospital with prevalence of serious illness 12% Derived prior to vaccination <i>Haemophilus influenzae</i> and pneumococcus
Pneumonia		
Crain et al., 1991 (238)	Cough, tachypnoea, rales, retractions, rhonchi, rhinorrhoea, wheezing	Derived in infants <2 months. Model not derived using multivariable statistical analysis
Meningitis		
Joffe et al., (239)	Suspicious findings on neurological and or physical examination, physician visit in previous 48 hours, convulsions on arrival at emergency department, a focal seizure	Outcome not verified in all children. Model not derived using multivariable statistical analysis
Offringa et al., 1992 (240)	Petechiae, nuchal rigidity, coma	Model not derived using multivariable statistical analysis
Pantell et al., 2004 (242)	Abnormal appearance, age, temperature	Derived in infants < 3 months
Oostenbrink et al., 2001 (241)	Duration of main complaint, vomiting, meningeal irritation, cyanosis, petechiae, disturbed consciousness	Derived in emergency department in children with meningeal signs and prevalence of serious infection 44%
Bacteraemia		
Schwartz & Wientzen, 1982 (243)	Appearance, appetite, cry, resistance to examination	Derived in children <36 months Derived prior to vaccination <i>Haemophilus influenzae</i> and pneumococcus Model not derived using multivariable statistical analysis

The prediction rules available at the time were thus judged to be unsuitable for the identification of serious bacterial infection in children presenting to primary care settings. It is acknowledged that methods have since been developed for updating existing rules to overcome differences between the derivation population and the population of interest (62, 247). However, methodological guidance on prediction rule updating was not available at the time this program of research was being conducted.

Given the limitations of the existing prediction rules and absence at the time of methodological guidance for adjusting prediction models to other settings, I proposed deriving

a prediction rule comprised of predictors from the patients' history and physical examination, using data sourced from primary care and low prevalence emergency department settings. The dataset to be used for this purpose was a prospective cohort study of 700 children aged 3 months to 16 years attending the Paediatric Assessment Unit at the University Hospital Coventry and Warwickshire NHS Trust in the United Kingdom with suspected acute infection. Just over half the children in this cohort (51%) were referred from primary care, 28.3% were self-referrals and the remainder (16.6%) emergency ambulance transfers. I then planned to validate this rule in multiple datasets being compiled by an international collaboration of researchers in the area. (222) In addition, simplified versions of the rule were to be derived and the accuracy and reclassification ability of the simplified prediction rules compared to the original prediction rule.

4.4.2 Determination of the diagnostic value of C-reactive protein (CRP)

Despite the widespread use of C-reactive protein (CRP) in some countries and its increasing uptake in others at the time (throughout 2006 and 2007), the diagnostic value of CRP for identifying serious infections in children was uncertain and had not been systematically studied. In order to inform of the ability of CRP to accurately identify serious bacterial infection in children at first presentation, and its added value beyond existing clinical information, two studies were proposed. The first was a systematic review of the accuracy and independent value of CRP for serious bacterial and bacterial infection in non-hospitalised children. The second study was a modelling study to determine the added value of CRP over and above patient history and clinical examination. In this study, the area under the receiver operating characteristic curve (AUROC) of the derived rule, comprised of predictors from the patients' history and physical examination (the base model), was to be compared to a model containing both the CRP test value and the derived rule (the extended model).

4.5 The research completed, incomplete and research plan modification

Of the four studies originally proposed to address the goal of this part of the thesis, two were not completed. An analysis of the dataset sourced to develop the prediction model was attempted but it was judged that a valid and clinically sensible prediction model comprised of variables from the medical history and physical examination could not be derived. Consequently, quantification of the added value of the inflammatory marker CRP beyond variables obtained from the medical history and physical examination could also not be determined.

Derivation of the prediction rule was judged to be unwise due to the frequency of missing predictor data and the limitations of available methods for handling missing data. Less than half of the cases in the dataset (47%) had complete data on all the predictors preselected for inclusion in statistical modelling, with predictor data missing for between 2% and 44% of cases. Serious bacterial infection was more common among cases with complete predictor data (17% vs 14% among cases with missing data) and a larger percent of cases with complete predictor data were considered to be urgent or very urgent on the Manchester Triage Scale (74% vs 70%). Several approaches to managing the missing data were considered including 1) removal of participants with missing predictor values, 2) removing predictors with missing data, 3) single imputation with the arithmetic mean or the predicted score from a regression equation and 4) multiple imputation. Removal of participants with missing predictor data was not considered an appropriate approach as it had been shown to lead to inaccurate estimates of the predictor outcome associations and the predictive performance of the final model. (248, 249) Removing predictors with missing data was considered, but judged to have an unacceptable trade-off in terms of the face value of the rule. The large amount of missing data for some predictors (e.g. respiratory rate) in the research setting suggested that predictor measurement may be difficult and it is likely that data for such variables will also be frequently missing in clinical practice. Removing these predictors from further analysis may be sensible in that it may lead to the derivation of a rule more likely to be applied in totality, with data from all the predictors used to derive the result. However, removal of such predictors would likely affect the face validity of the rule and threaten clinicians' acceptance and implementation of it. Single imputation, while considered to be preferable to the deletion methods described, had been shown to produce biased parameter estimates and or to attenuate standard errors. (249) The final approach, multiple imputation, was considered to be the optimal method. Based on simulation studies and theoretical reasoning at the time, multiple imputation was assumed to yield unbiased results and appropriate standard errors. (250)

Today, multiple imputation is accepted and advocated as the preferred approach for managing missing data in studies deriving multivariable models and is embedded in commonly used statistical packages. (1) At the time this study was undertaken however, the methods of multiple imputation were in development. (251) Definitive guidelines on the allowable proportion of missing data to which multiple imputation techniques could be validly applied were not available, and there were limited studies investigating the effects of varying amounts of missing values, the mechanisms of missingness, or approaches to handling missing data, on model performance to guide the analysis. Due to the large amount of missing data, limitations

of conventional methods for replacing missing values, and concern regarding the clinical sensibility and validity of a prediction rule derived using these methods, derivation of the proposed prediction rule was not completed.

The remaining two studies were completed; one as proposed, and the second with some modification to the protocol. The first of these was a systematic review to determine the accuracy and independent value of C-reactive protein for the recognition of serious bacterial infection relative to the prevailing reference standard. This study is presented in the following chapter. The second completed study determined the effect of simplification of a prediction rule on performance (Chapter 6). When the planned prediction rule for identifying serious bacterial infection in children in a primary care setting could not be derived, an alternative dataset was sourced to allow the study to be conducted. Though not in the clinical area of interest, the issues raised are pertinent to diagnostic prediction rules developed for any clinical setting.

4.6 The current status of diagnostic prediction rules as a strategy for assisting clinicians' diagnostic management of children with possible serious bacterial infection presenting to primary care

The diagnostic management of children with possible serious bacterial infection has continued to be an area of intense interest among researchers and policy makers. (252, 253) In recent years several clinical prediction rules for identifying serious infection in children presenting to the emergency department and primary care have been derived and evaluated (220, 254, 255) and some of the clinical prediction rules derived prior to the end of 2007 and described in Table 4.1, have undergone further development. (236, 237) The characteristics of, and stage of development of these clinical prediction rules which are comprised of clinical predictors or clinical and laboratory predictors such as C-reactive protein are presented in Table 4.2.

Table 4.2 Clinical prediction rules comprised of predictors from the history or physical examination (with or without laboratory predictors) for serious bacterial infection overall or specific serious bacterial infections in children derived and or evaluated since 2007

Clinical prediction rule/Setting in which rule derived	Predictors included in the rule	Stage of development*	Findings
All serious infections			
Bleeker et al., 2001 (236)/ED	Clinical	Narrow validation (Bleeker et al., 2007 (256))	Discriminative ability of the prediction rule was poor in the validation dataset (AUROC derivation 0.75 95%CI 0.68 to 0.83 versus AUROC validation 0.60 95%CI 0.49 to 0.70). After updating the rule the AUROC was 0.69 95%CI 0.63 to 0.75
		Impact study of randomised controlled design (Roukema et al., 2008 (198))	Increased length of stay in the emergency department [†] and number of laboratory tests ordered with implementation of the clinical prediction rule
Bleeker et al., 2001 (236)/ED	Clinical + laboratory	Narrow validation (Bleeker et al., 2007 (256))	Discriminative ability of the prediction rule was similar in the validation dataset (AUROC derivation 0.83 95%CI 0.77 to 0.89 versus AUROC validation 0.78 95%CI 0.69 to 0.86)
Yale observation scale McCarthy et al., 1982 (237)/ED	Clinical	Broad validation (NICE, 2007 (253))	Sensitivity was low in low prevalence settings at different cut-offs (46.2% and 23.1%), intermediate prevalence settings at different cut-offs (40.5% and 22.3%) and high prevalence settings at different cut-offs (30.3% and 19.5%)
Four-step decision tree (255)/PC	Clinical	Internal validation Broad validation (NICE, 2007 (253))	Sensitivity 96.8 (95%CI 83.3 to 99.9) Sensitivity 90% in one low prevalence setting, ranged from 75.5% to 87.8% in two intermediate prevalence settings and from 23.0% to 89.1% in two high prevalence settings
		Narrow validation (Verbakel et al., 2015 (257))	Sensitivity 100% (95%CI 71.5% to 100%) in the general practice setting, 82.7% (95%CI 72.2% to 92.4%) in pediatric outpatient setting and 69.5% (95%CI 62.6% to 75.9%) in the emergency department setting
Nijman et al., 2013 [‡] (254)/ED	Clinical + laboratory	Narrow validation (Nijman et al., 2013 (254)) Impact study of randomised controlled design (de Vos-Kerkhof et al., 2015 (204))	C statistic§ 0.69 (95%CI 0.53 to 0.86) in validation dataset versus 0.86 (95%CI 0.79 to 0.92) in the derivation dataset No difference in 'correct' diagnoses (false positives and false negatives) between study groups [†] . No difference in median length of stay. In the intervention group fewer full blood counts were performed and more urine dipsticks were correctly done according to current guidelines.
Pneumonia			
Nijman et al., 2013 (254)/ED	Clinical + laboratory	Narrow validation (Nijman et al., 2013 (254))	C statistic 0.81 (95%CI 0.69 to 0.93) in validation dataset versus 0.80 (95%CI 0.72 to 0.89)
		Impact study of randomised controlled design (de Vos - Kerkhof et al., 2015 (204))	No difference in 'correct' diagnoses (false positives and false negatives) between study groups [†] . No difference in median length of stay. In the intervention group fewer full blood counts were performed and more urine dipsticks were done according to current guidelines.

Table 4.2 Continued**Pneumonia**

Craig et al., 2010 (220)/ED	Clinical	Narrow validation (Craig et al., 2010 (220))	AUROC 0.84 (95%CI 0.82 to 0.87) in the validation dataset versus 0.84 (95%CI 0.83 to 0.86) in derivation dataset
Pneumonia rule (255)/PC	Clinical	Internal validation (Van den Bruel et al., 2007 (255))	Sensitivity was 93.8% (95%CI 69.8 to 99.8) for the predictor variables 'dyspnoea' and 'something is wrong'. Sensitivity was 93.8% (95%CI 69.8 to 99.8) for the predictor variables 'dyspnoea' and 'different illness'
		Broad validation (Nice, 2007 (253))	Sensitivity ranged from 92.4% to 94.1% in two low prevalence settings and from 26.9% to 81.5% in three intermediate prevalence settings
Oostenbrink., 2013 (258)/ED	Clinical + laboratory	Broad validation (Oostenbrink et al., (258))	C statistic 0.86 and 0.86 in two validation datasets
Bacteraemia			
Craig et al., 2010 (220)/ED	Clinical	Narrow validation (Craig et al., 2010 (220))	AUROC 0.74 (95%CI 0.66 to 0.82) in the validation dataset vs AUROC 0.88 (95%CI NA) in the derivation dataset
Meningitis			
Meningitis rule (240)/ED	Clinical	Broad validation (NICE, 2010 (253))	Sensitivity ranged from 33.3% to 100% in two low prevalence settings, and was 95.5% in one high prevalence setting

ED – emergency department; AUROC – area under the receiver operator characteristic curve; PC – primary care; NA – not available

* Narrow validation is application of the rule in a similar clinical setting and population to derivation data. Broad validation is application of the rule in multiple settings and varying prevalence of disease; † primary outcome; ‡ pneumonia not included; § c statistic equivalent to the AUROC

Review of Table 4.2 indicates that efforts to develop a clinical prediction rule with demonstrated benefit when applied in practice have, as yet, been both incomplete and unsuccessful. This is particularly the case for primary care settings. Despite the need for formal impact analysis to assess whether care provided when clinicians have access to a clinical prediction rule either leads to better outcomes for the patient, or provides other benefits relative to the current diagnostic pathway without adversely affecting patients, only two clinical prediction rules (one comprised only of clinical variables, and the other of clinical variables and the laboratory test C-reactive protein) have had the effects of application in practice assessed, and these were in emergency department settings. (198, 204) In these two studies, exposure to the clinical prediction rule did not lead to improvements in process of care outcomes. In the first of these studies, care provided with a diagnostic prediction rule had no effect on 'correct' diagnoses (primary outcome) or length of stay in the emergency department, (204) and in the second, care provided with a diagnostic prediction rule did not shorten length of stay in the emergency department and increased the number of laboratory tests ordered. (198) Neither study reported patient outcomes.

In addition, to date there have been few studies deriving or validating clinical prediction rules for identifying serious infection in children in the primary care setting (Table 4.2), and though the one rule derived in this setting, the four-stage decision tree, has demonstrated value for ruling out serious infection, residual uncertainty remains. (255) The four-step decision tree developed in a general practice population in Belgium had high sensitivity (90%) when validated in one low prevalence setting, and 100% sensitivity in a general practice setting similar to the setting in which it was derived. (257) However, false positives are common (specificities of 44% and 77% in validation studies) and consequently additional clinical assessment or testing may be needed to reduce unnecessary referrals. Clinical prediction rules specifically for pneumonia have also shown value for ruling out pneumonia in some validation studies in low prevalence settings, (sensitivities of 92.5% and 94.1%), but again, the percentage of false positives was high (specificities <45%).

At this point in time, further work is needed before diagnostic clinical prediction rules based on clinical features for identifying serious infection in children presenting to primary care be implemented in practice. Extensive prospective validation in the primary care setting and updating of existing rules should be undertaken in preference to the derivation of new rules, with the performance of the prediction rule compared to the existing diagnostic strategy. Finally, if clinical prediction rules with good performance in validation studies are identified, studies of the impact of application of the prediction rule, ideally reporting relevant patient outcomes, are an essential final step for determining whether the clinical prediction rule does more good than harm.

Chapter 5 The diagnostic accuracy and independent value of C-reactive protein for detecting bacterial infection in nonhospitalised infants and children with fever

This is an accepted manuscript of an article published by Elsevier in the Journal of Pediatrics on the 27th of May 2008, available online: doi: 10.1016/j.jpeds.2008.04.023. © 2008.

Sanders S, Barnett A, Correa-Velez I, Coulthard M, Doust J. 2008 Systematic review of the diagnostic accuracy of C-reactive protein to detect bacterial infection in hospitalized infants and children with fever. *Journal of Pediatrics*. 153(4):570-4, Elsevier, 2008.

This manuscript version is made available under the CC-BY-NC-ND 4.0 licence

[http://creativecommons.org/licenses/by-nc-](http://creativecommons.org/licenses/by-nc-nd/4.0/)

[nc/4.0/\(http://creativecommons.org/licenses/by-nc-nd/4.0/\)](http://creativecommons.org/licenses/by-nc-nd/4.0/)

5.1 Preface to Chapter 5

C-reactive protein is an inflammatory biomarker that is now widely utilised as a laboratory and point of care test for assisting in the diagnostic management of children with possible serious bacterial infection. At the time this program of research commenced, however, the accuracy of C-reactive protein for detecting bacterial infection in children presenting to primary care had not been established. This chapter presents the results of a systematic review conducted to determine the accuracy and independent value of the-reactive protein for detecting bacterial infection in non-hospitalised infants and children.

5.2 Abstract

Objective: To determine the accuracy of C-reactive protein for diagnosing serious bacterial and bacterial infections in infants and children presenting with fever.

Study design: Systematic review of diagnostic accuracy studies. We included studies comparing the diagnostic accuracy of C-reactive protein with microbiologic confirmation of (a) serious bacterial and (b) bacterial infection.

Results: For differentiating between serious bacterial infection and benign or nonbacterial infection (6 studies), the pooled estimate of sensitivity was 0.77 (95% CI 0.68 to 0.83); specificity 0.79 (95% CI 0.74 to 0.83); positive likelihood ratio, 3.64 (95% CI 2.99 to 4.43); and negative likelihood ratio, 0.29 (95% CI 0.22 to 0.40). In multivariate analysis, C-reactive protein is an independent predictor of serious bacterial infection. 3 studies investigating the accuracy of C-reactive for diagnosing bacterial infection could not be pooled, but all showed a lower sensitivity compared with studies using serious bacterial infection as the reference diagnosis.

Conclusions: C-reactive protein provides moderate and independent information for both ruling in and ruling out serious bacterial infection in children with fever at first presentation. Poor sensitivity means that C-reactive protein cannot be used to exclude all bacterial infection.

5.3 Introduction

Although fever is a common reason for children to be brought to medical attention, it remains a diagnostic and management challenge. Most children with fever will have a self-limiting illness that resolves in a few days without active intervention. Some children, however, benefit from antibiotic treatment, and some substantially so. Few clinical features distinguish those who benefit from those who do not, and doctors' clinical judgment frequently is not able to distinguish children with bacterial infection from non-bacterial infection. (259) Because both doctors and patients are aware of the potential benefits of antibiotics for some patients and the potential risk of not treating a life-threatening infection in a small minority, there is an incentive to prescribe antibiotics, despite the costs, adverse effects, and increasing antibiotic resistance that result from this practice. (260-262)

C-reactive protein (CRP) rises in response to infectious and inflammatory diseases and shows greater elevations in serious bacterial than in other bacterial infections. It may distinguish those children who have a bacterial infection that could benefit from antibiotics from those who do not. (234, 263, 264) Our aim in this review was to evaluate the diagnostic accuracy of CRP in infants and children with an initial complaint of fever (with or without signs of respiratory tract infection).

5.4 Methods

Because clinical assessment and prior diagnostic testing may change the spectrum of patients being assessed and therefore the diagnostic accuracy of a test, (265) we limited this review to studies conducted in children who came to medical attention initially with a complaint of fever. Studies that included children who had been admitted to hospital (other than in an emergency department observation ward) were excluded from the review.

We included studies that compared a blood or serum CRP measurement with a reference standard of a microbiologic diagnosis of (a) serious bacterial infection (versus benign bacterial or nonbacterial infection) or (b) bacterial infection (versus nonbacterial infection). We excluded studies in which more than 10% of participants were neonates, and the reference standard was for the diagnosis of a single specific disease (e.g. meningitis, gastroenteritis, or arthritis) or studies conducted in subgroups of patients with specific medical conditions, such as cancer or renal failure. No language restriction was applied. Ethics approval was not required to conduct this study.

5.4.1 Identification of studies

We searched the databases Medline and EMBASE from inception to December 2007 using the following terms: C-reactive protein (MeSH) OR C-reactive protein (text word) OR CRP (text word) AND bacterial infections (MeSH) OR virus disease (MeSH) OR bacteria* (text word) OR virus (text word) OR viral (text word) AND (child (MeSH) OR child* (text word) OR infant (MeSH) OR infant* (text word)). We checked the reference lists of all included papers and review articles and forward searched any identified papers. 2 reviewers screened the titles and abstracts from the electronic searches against the inclusion and exclusion criteria. Disagreements were resolved by discussion with a third reviewer.

5.4.2 Quality assessment

Quality of the included studies was assessed independently by 2 reviewers using QUADAS, (266) a validated tool for assessing quality of diagnostic studies. We used the 11-item version as recommended by the Cochrane Diagnostic Test Accuracy Working Group, which contains items relating to patient spectrum, reference standard, disease progression bias, verification bias, review bias, incorporation bias, test execution, study withdrawals, and indeterminate results. 2 of the 11 items were deemed not relevant due to the objectivity of the CRP test and were omitted. 1 item that asks whether patients received the same reference standard regardless of the index test result was split into 2 because it was possible that a different reference standard was applied but performance of the reference test was not related to the outcome of the index test. Percentage agreement and the κ statistic were calculated to assess the interobserver variation of the initial assessment of both reviewers.

5.4.3 Data extraction

Data were extracted by 2 reviewers independently on predesigned and piloted forms. Where necessary, we contacted authors for data or clarifying information.

5.4.4 Data analysis

Data from each study were extracted in 2 x 2 tables. We combined the categories of invasive bacterial and localised bacterial infection in the category of bacterial infection for the purpose of the 2 x 2 tables; similarly, we categorised mixed bacterial and viral infections (1 study) and proven or possible bacterial infections (1 study) as bacterial. In a sensitivity analysis, these categorisations had a negligible effect on the results. Heterogeneity in study results was examined graphically using plots of sensitivity versus 1 minus specificity. Clinical heterogeneity was examined using descriptions of study characteristics. Where there was sufficient homogeneity of results, we used a random-effects bivariate model to obtain summary

estimates of sensitivity and specificity and corresponding 95% confidence intervals. This model estimates and incorporates the correlation that may exist because of the trade-off between sensitivity and specificity due to changes in the threshold in the index test for defining disease used between studies. It is equivalent to a hierarchical summary receiver operating characteristic curve analysis in most situations. (267) The model also estimates the diagnostic odds ratio, a measure of overall test accuracy equivalent to the ratio of true to false test results. (268) Statistical codes were kindly provided by Roger Harbord of Bristol University. (267) Stata 9 was used for all analysis.

5.5 Results

5.5.1 Study characteristics

The search retrieved 1770 potentially relevant titles and abstracts. Of these, 10 studies assessing a total of 2046 participants met the inclusion criteria for the review. All of the studies were conducted in emergency departments. 36 studies examining the diagnostic accuracy of CRP in children admitted to hospital were excluded. Characteristics of the studies investigating CRP for the identification of serious bacterial infection and for the differentiation of bacterial and nonbacterial/viral infection are summarised in Table 5.1.

One study (236) used chart review to identify children with fever and no localising signs as part of a larger prospective study; all other studies were prospective. Fever was not defined in 4 studies. (269-272) 4 studies (244, 245, 273, 274) included a small proportion of participants (less than 10%) younger than 1 month of age, and in 1 study (271) the minimum age was not reported and could not be obtained from the author.

5.5.2 Quality assessment

Result of the assessment of quality using the QUADAS checklist for the 6 studies with 2 x 2 data are shown in Table 5.2. In almost all studies, withdrawals and handling of uninterpretable results were poorly reported. None of the studies provided sufficient detail to judge whether those interpreting the results of the reference standard tests were blind to the CRP results. The κ statistic for interobserver variation in the initial quality assessment, before discussion with the third reviewer, was 0.53.

Table 5.1. Details of studies investigating the accuracy of C-reactive protein (CRP) to differentiate serious bacterial infection from benign bacterial/nonbacterial infection and bacterial from nonbacterial/viral infection

Study, year published	Population			Index test	
	Setting: Location	N	Inclusion criteria	Mean age (range)	Sample: Assay type
Studies investigating CRP for the differentiation of serious bacterial infection from benign bacteria/nonbacterial infection and providing data for construction of a 2x2 table					
Pulliam et al., 2001 (272)	ED: USA	77	Fever (undefined) without apparent source after history and examination. Excluded children with AOM, pharyngitis, clinical pneumonia, acute RTI, acute gastroenteritis	9.7 months (1 to 35 months)	Serum: Turbidometric immunoassay
Lacour et al., 2001 (244)	ED: Switzerland	124	Temperature >38°C without localising signs of infection in history or on physical examination	10.9 and 11.2 months (7 days to 36 months)	Whole blood: rapid immunometric method*
Isaacman and Burke, 2002 (270)	ED: USA	256	Fever (undefined) without apparent source	NR (3 to 36 months)	Serum: heterogeneous immunoassay format
Galetto-Lacour et al., 2003 (245)	ED: Switzerland	99	Temperature >38°C without localising signs of infection in history or on physical examination	NR (7 days to 36 months)	Whole blood: rapid immunometric method*
Berger et al., 1996 (274)	ED: Netherlands	127	Rectal temperature ≥38°C	NR (14 days to 1 year)	NR: NR
Andreola et al., 2007 (273)	ED: Italy	408	Fever (rectal temperature >38°C) of uncertain source	Median 10 months (2 to 16 months)	Whole blood: nephelometric assay
Studies using multivariable modelling to determine the predictive value of CRP for the detection of serious bacterial infection					
Bleeker et al., 2001 (236)	ED: Netherlands	231	Fever ≥38°C without apparent source	13 months (1 to 36 months)	Serum: NR
Pulliam et al., 2001† (272)	As above				
Isaacman and Burke, 2002† (270)	As above				
Berger et al., 1996† (274)	As above				
Andreola et al., 2007† (273)	As above				
Studies investigating CRP for the differentiation of bacterial infection from nonbacterial/viral infection					
McCarthy et al., 1978 (271)	ED: USA	400	Febrile (undefined). May or may not have localising signs	43.8 months (NR)	Serum: NR
Tejani et al., 1995 (275)	Participants in RCT: USA	185	Included children with signs and symptoms of acute otitis media (symptoms, tympanic membrane signs, MEF on tympanocentesis)	27 months (3 months to 7 years)	Serum: Rate nephelometry
Cobben et al., 1990 (269)	ED: Netherlands	139	Symptoms of infection (including fever, vomiting, cough, dyspnoea, stridor, abdominal cramping, excluding AOM)		

ED – emergency department; AOM – acute otitis media; RTI – respiratory tract infection; NR – not reported; BC – blood culture; UC – urine culture; FBE – full blood examination; UA – dipstick analysis; STCU – stool culture; ENT – ear, nose or throat swab culture; CXR – chest x-ray; ESR – erythrocyte sedimentation rate; RCT – randomised controlled trial; MEF – middle ear fluid + test conducted in all children; - test not conducted or not reported; ± test conducted in some children

*15 minutes from test to result; †Study provided data for construction of a 2x2 table and performed multivariable modelling

Table 5.1. Continued

Study, year published	Reference tests; methods used to detect the cause of infection											Outcome definition
	FBE	ESR	UA or AC	CXR	CSF	BC	STCU	ENTC	F-up	Other tests	Viral testing	
Studies investigating CRP for the differentiation of serious bacterial infection from benign bacteria/nonbacterial infection and providing data for construction of a 2x2 table												
Pulliam et al., 2001 (272)	+	+	+	±	±	+	-	-	-	-	-	Proven serious bacterial infection (positive blood or urine culture or local infiltrate on CXR)
Lacour et al., 2001 (244)	+	-	+	±	±	±	-	-	+	-	-	Serious bacterial infection (positive urine, blood or CSF culture, or local infiltrate on CXR)
Isaacman and Burke, 2002 (270)	+	-	±	±	-	+	-	-	-	-	-	Occult bacterial infection (positive blood or urine culture or local infiltrate on CXR)
Galetto-Lacour et al., 2003 (245)	+	-	±	-	±	±	-	-	+	-	-	Serious bacterial infection (positive blood, urine, CSF culture, local infiltrate on CXR, positive CT and surgical exploration for deep abscess)
Berger et al., 1996 (274)	+	+	+	-	-	+	+	+	+	±	-	Serious bacterial infection (positive blood, urine, CSF, stool or specimen culture, or pulmonary infiltrate on CXR)
Andreola et al., 2007 (273)	+	+	+	±	±	±	±	±	±	±	-	Serious bacterial infection (positive blood, urine, CSF culture, infiltrate on CXR, sepsis according to criteria of signs and symptoms of inflammation plus infection, tachycardia, decreased capillary refill or mottling and 1 indication of altered organ function)
Studies using multivariable modelling to determine the predictive value of CRP for the detection of serious bacterial infection												
Bleeker et al., 2001 (236)	-	-	±	-	±	±	±	-	+	±	-	Serious bacterial infection (positive culture or consensus diagnosis and negative follow-up)
Pulliam et al., 2001† (272)	As above											
Isaacman and Burke, 2002† (270)	As above											
Berger et al., 1996† (274)	As above											
Andreola et al., 2007† (273)	As above											
Studies investigating CRP for the differentiation of bacterial infection from nonbacterial/viral infection												
McCarthy et al., 1978 (271)	±	±	±	+	±	±	±	-	-	±	±	Positive blood or other culture
Tejani et al., 1995 (275)	+	-	-	-	-	-	-	+	+	-	+	Positive culture in MEF or nasal wash specimens (including mixed bacterial or viral infections)
Cobben et al., 1990 (269)						NR						Positive culture, repeat clinical evaluation, infiltrate on CXR, recovery with antibiotics (including proven or possible bacterial infection)

ED – emergency department; AOM – acute otitis media; RTI – respiratory tract infection; NR – not reported; BC – blood culture; UC – urine culture; FBE – full blood examination; UA – dipstick analysis; STCU – stool culture; ENTC – ear, nose or throat swab culture; CXR – chest x-ray; ESR – erythrocyte sedimentation rate; RCT – randomised controlled trial; MEF – middle ear fluid
 + test conducted in all children; - test not conducted or not reported; ± test conducted in some children

*15 minutes from test to result; †Study provided data for construction of a 2x2 table and performed multivariable modelling

Table 5.2. Quality of the included studies with useable 2 x 2 data according to QUADAS criteria

Study (year published)	1. Was the spectrum of patients representative of the patients who will receive the test in practice?	2. Is the reference standard likely to correctly classify the target condition?	3. Is the time period between reference standard and index test short enough to be reasonably sure the target condition did not change between the 2 tests?	4. Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?	5. Did patients receive the same reference test?	5A. Was the reference standard conducted regardless of the index test result?	6. Was the reference standard independent of the index test?	7. Were the reference standard results interpreted without knowledge of the results of the index test?	8. Were uninterpretable/intermediate test results reported?	9. Were withdrawals from the study explained?
Children with clinically undetectable source of acute fever (specifically excludes children with localising signs)										
Pulliam et al., 2001 (272)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Unclear	Unclear
Lacour et al., 2001 (244)	Yes	Yes	Yes	Yes	No	Yes	No	Unclear	Unclear	Yes
Isaacman and Burke, 2002 (270)	No	Yes	Yes	Yes	No	Yes	Yes	Unclear	Unclear	Unclear
Galetto-Lacour et al., 2003 (245)	Yes	Yes	Yes	Yes	No	Yes	Yes	Unclear	Yes	Yes
Andreola et al., 2007 (273)	Yes	Yes	Yes	Yes	No	Yes	Yes	Unclear	Yes	Yes
Children with acute fever (may or may not have localising signs)										
Berger et al., 1996 (274)	Yes	Yes	Unclear	Yes	No	Yes	Yes	Unclear	Yes	Yes
McCarthy et al., 1978 (271)	Unclear	Unclear	Unclear	Yes	No	Yes	Yes	Unclear	Unclear	Unclear
Children with signs or symptoms of respiratory tract infection										
Tejani et al., 1995 (275)	No	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Unclear	Unclear
Symptoms of infection										
Cobbien et al., 1990 (269)	Yes	Unclear	Unclear	Yes	Yes	Yes	Yes	Unclear	Unclear	Unclear

*1 QUADAS item has been modified (item 5) and 2 QUADAS items have been omitted.

5.5.3 CRP for the detection of serious bacterial infection

7 of the 10 studies included in the review (1090 children) investigated the accuracy of CRP to differentiate serious bacterial infection from self-limiting bacterial or nonbacterial infection. (236, 244, 245, 270, 272-274) 1 of these 7 studies did not provide data for constructing a 2x2 table. (236) Five studies conducted multivariate modelling to determine the independent predictive value of CRP. (236, 270, 272-274) The definition of serious bacterial infection varied somewhat but all included evidence of a bacterial infection on blood culture, chest radiograph, lumbar puncture, or urine culture. (Table 5.1) The prevalence of serious bacterial infection ranged from 11% to 29%. Visual inspection of the data demonstrated sufficient homogeneity to allow summary estimates of sensitivity and specificity. Summary estimates from the bivariate model were sensitivity, 0.77 (95% CI 0.68 to 0.83); specificity 0.79 (95% CI 0.74 to 0.83); and diagnostic odds ratio 12.29 (95% CI 8.51 to 17.75) (Figure 5.1). The pooled positive likelihood ratio was 3.64 (95% CI 2.99 to 4.43), representing a small change in the probability of serious bacterial infection with a positive CRP test (Table 5.3). The pooled negative likelihood ratio was 0.29 (95% CI 0.22 to 0.40), suggesting a moderate change in the probability of a serious bacterial infection with a negative CRP.

In the 5 studies performing multivariate analysis (236, 270, 272-274), CRP was an independent predictor of serious bacterial infection, that is, CRP adds information to other clinical features in determining the presence or absence of serious bacterial infection.

5.5.4 CRP for the detection of bacterial and viral infection

3 studies (722 children) assessed the accuracy of CRP to differentiate bacterial from viral infection (Table 5.3). (269, 271, 275) The prevalence of bacterial infection was 28%, 35% and 82%, the latter occurring in a study that included a category of mixed and bacterial infections. Given the significant heterogeneity of these studies, we did not estimate pooled summaries of sensitivity and specificity. However, each of these studies showed more limited sensitivity than did the studies assessing CRP for the diagnosis of serious bacterial infection.

Figure 5.1. Results of studies estimating the sensitivity and specificity of C-reactive protein for the detection of serious bacterial infection and bacterial infection.

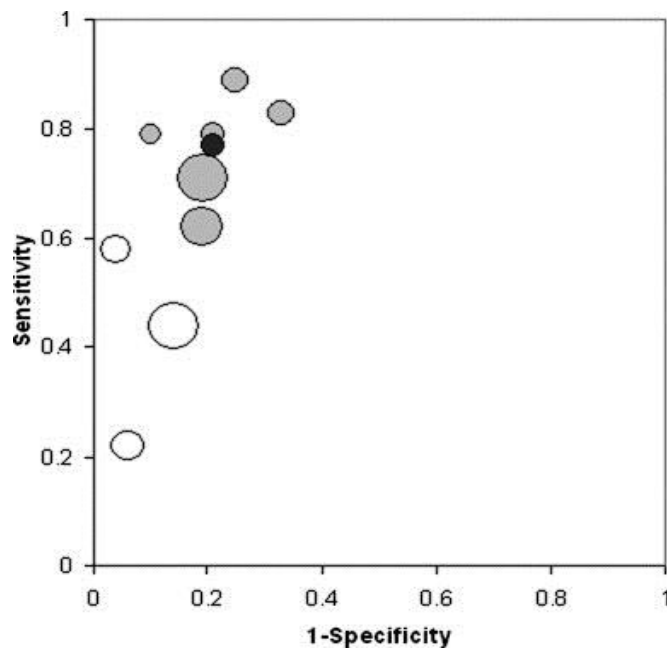


Figure interpretation: Grey circles indicate serious bacterial infection; white circles indicate bacterial infection; the black circle indicates summary point for serious bacterial infection (summary estimate from the random-effects bivariate analysis). Each circle represents the sensitivity and specificity of the individual studies included in the review. The size of the circle is proportionate to study size.

Table 5.3. Performance of C-reactive protein (CRP) for the detection of serious bacterial infection and bacterial infection

Study, year published	CRP cut-off	Prevalence	TP	FN	TN	FP	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)
CRP for the detection of serious bacterial infection*										
Isaacman and Burke, 2002 (270)	4.4 mg/dL	11%	18	11	184	43	62 (44 to 80)	81 (76 to 86)	3.28 (2.22 to 4.85)	0.47 (0.29 to 0.75)
Pulliam et al, 2001 (272)	7.0 mg/dL	18%	11	3	57	6	79 (57 to 100)	90 (83 to 98)	8.25 (3.68 to 18.52)	0.24 (0.09 to 0.65)
Lacour et al, 2001 (244)	40 mg/L	23%	25	3	72	24	89 (78 to 100)	75 (66 to 84)	3.51 (2.47 to 5.17)	0.15 (0.05 to 0.42)
Galletto-Lacour et al, 2003 (245)	40 mg/L	29%	23	6	55	15	79 (65 to 94)	79 (69 to 88)	3.70 (2.28 to 6.02)	0.26 (0.13 to 0.54)
Berger et al, 1996† (274)	20 mg/L	24%	25	5	65	32	83 (70 to 97)	67 (58 to 76)	2.53 (1.82 to 3.49)	0.25 (0.11 to 0.56)
Andreola et al, 2007† (273)	>40 mg/L	23%	67	27	258	56	71 (61 to 80)	81 (76 to 85)	3.79 (3.05 to 5.03)	0.35 (0.25 to 0.48)
Pooled likelihood ratios										
									3.64 (2.99 to 4.43)	0.29 (0.22 to 0.40)
CRP for the detection of bacterial infection										
McCarthy et al, 1978†§ (271)	Positive†	28%	48	60	250	40	44(35 to 54)	86 (82 to 90)	3.22 (2.26 to 4.60)	0.64 (0.54 to 0.77)
Tejaniet al, 1995 (275)	>2mg/dL	82%	33	118	32	2	22 (15 to 28)	94 (86 to 100)	3.72 (0.94 to 14.75)	0.83 (0.74 to 0.94)
Cobben et al, 1990 (269)	35 mg/L	35%	28	20	87	4	58 (44 to 72)	96 (91 to 100)	13.3 (4.94 to 36.0)	0.44 (0.31 to 0.61)

* Results of 6 of 7 studies investigating the accuracy of CRP for detecting serious bacterial infection that provide data for the construction of a 2x2 table; †Some children <1 month of age; ‡Positive test indicates level of >6 mg/L or >10 mg/L, depending on the test kit or reagent used.

§Deep, superficial, and possible bacterial combined versus nonbacterial infection; || Bacterial only and mixed viral bacterial versus nonbacterial.

5.6 Discussion

We conducted a systematic review of studies assessing the diagnostic accuracy of CRP for the diagnosis of bacterial infections in children initially evaluated because of fever. The results indicate that CRP is of moderate value for ruling out serious bacterial infection in a child with a fever but is of limited value for ruling out all bacterial infections. The diagnostic accuracy of the test for all bacterial infections is limited by the significant overlap in CRP values for children with viral and bacterial infections, especially in the early stages of a bacterial infection, (234, 276-278) and the significant interindividual variation in test values. (279)

All of the studies included in this review were conducted in emergency departments. This may limit the applicability of the results of the review for other setting in which children present with fever, such as general practice and outpatient pediatric clinics. The prevalence of bacterial infections seen in the studies included in this review was likely to be higher than would be seen in general practice. (280) As an indication of how the change in the pre-test probability might influence the clinical utility of the test, if we assume the prevalence of serious bacterial infection among children presenting with fever to an emergency department is 7%, (281) the probability of serious bacterial infection given a positive CRP test is estimated to be 22% and given a negative test is 2%. Assuming a pre-test probability of serious bacterial infection of 1% in general practice, (219) the probability of serious bacterial infection given a positive CRP test is estimated to be 4% and given a negative test is 0.3%.

An earlier review in a more limited set of studies concluded that there was no evidence to support the use of CRP in children with fever or to make decisions regarding the initiation or suspension of antibiotics. (282) Our review, however, indicates that CRP contributes moderate independent information. Our study has included more recent studies and excluded 4 studies included in the prior review that were conducted in children admitted to hospital. A systematic review of the diagnostic accuracy of CRP for bacterial infection in adults concluded that it was not sufficiently sensitive or specific. (283)

Procalcitonin is a biomarker that has been developed more recently and may also be useful in this patient group. In this review, 3 studies (244, 245, 273) compared measurement of procalcitonin and CRP, with 2 showing that procalcitonin had higher sensitivity and negative predictive value than CRP (93% vs 89% and 93% vs 79%). In the third study, CRP had a higher sensitivity and negative predictive value at the optimum cut-off values (69% vs 84%).

The results of this review indicate that CRP provides moderate but independent information in ruling in and ruling out serious bacterial infection in febrile children at first presentation. The test needs to be considered in the context of the other clinical findings and should not be relied on to excluded bacterial infection.

5.6.1 Acknowledgments

The authors wish to acknowledge the contribution made by Roger Harbord who provided the STATA ado file to complete the data analysis for this paper.

5.6.2 Author contributions

Conceived and designed the study: SS, JD, AB. Analysed the data: SS, IC-V, JD. Wrote the first draft of the manuscript: SS. Contributed to writing of the manuscript: SS, AB, IC-V, MC, JD. Agree with manuscript results and conclusions: SS, AB, IC-V, MC, JD. International Committee of Medical Journal Editors (ICMJE) criteria for authorship read and met: SS, AB, IC-V, MC, JD.

SS received funding from an Australian Postgraduate Award Scholarship. JD position was supported by a University of Queensland New Staff Start-up Grant. The funding source had no involvement in the conduct of this study.

5.7 The current status of C-reactive protein for assisting in the diagnostic management of non-hospitalised children with fever and possible serious infection

Since publication of this review in 2008, considerable research has been undertaken with the aim of clarifying the diagnostic accuracy of C-reactive protein and determining its value in clinical practice. In a more broadly focused but comparable systematic review funded by the Health Technology Assessment Program (HTA) in the United Kingdom, researchers determined the value of all possible blood tests, including C-reactive protein, for ruling in and ruling out serious infection in children in ambulatory settings. (253) The review included 5 studies of C-reactive protein published to mid-2009 and discussed the systematic review presented in this chapter. Three of the five studies included in the HTA review were included in the review presented in this chapter, and two studies included in the HTA review were excluded from the review presented in this chapter as they did not meet the reviews inclusion criteria. Despite this difference in included studies, the conclusions of both reviews with regard to the diagnostic accuracy of C-reactive protein were similar. The HTA review reported a pooled positive LR of 3.64 (95%CI 2.99 to 4.43) compared to a positive LR of 3.15 (95%CI 2.67 to 3.71) in the review presented in this chapter and a negative LR of 0.29 (95%CI 0.22 to 0.40) and 0.33 (95%CI 0.22 to 0.49) respectively. The HTA review also considered studies directly comparing

laboratory tests, and examined test cut-off values, finding that both C-reactive protein and procalcitonin offer similar diagnostic performance and are superior to white blood cell count. However, neither C-reactive protein nor procalcitonin had sufficient value to confirm or exclude a serious infection and different cut-off values were necessary depending on the purpose of the test (rule in or rule out) for the particular setting. (253)

More recently, research on C-reactive protein has been undertaken to determine the value of the test beyond its diagnostic accuracy. In a recent prospective observational study of febrile children presenting to the emergency department, introduction of bedside C-reactive protein testing substantially lowered length of stay in the emergency department. (284) Though not reporting on studies in children specifically, a recent Cochrane systematic review including six randomised controlled trials found that point of care C-reactive protein test used as an adjunct to primary care clinicians' clinical examination reduced antibiotic prescribing amongst individuals with acute respiratory infections. However, C-reactive protein did not affect patient reported outcomes compared to standard care. (285) Understanding of the value of C-reactive protein in the workup of children with fever will likely be advanced when the results of a currently in process two part clinical trial become available. The ERNIE2 studies are aimed at evaluating 1) the added value of point of care C-reactive protein in children with acute illness and considered on the basis of a clinical prediction rule to have a potentially serious illness (part A) (286), and 2) the effect of a multifaceted intervention comprising a point of care C-reactive protein test and /or safety netting advice in children consulting a primary care clinician, and who, on the basis of a clinical prediction rule are determined to not have a potentially serious illness (part B). (287) The primary outcome in this cluster randomised factorial controlled trial, and the outcome for which this trial is powered to detect a difference, is clinicians' immediate antibiotic prescribing rate with clinical recovery a secondary endpoint. The randomised component of the ERNIE 2 study , when complete will likely provide the best available advice to inform use of C-reactive protein in children in primary care practice , though limitations of patient management as an intermediate step in the pathway to affecting patient outcomes must be acknowledged. (288)

Chapter 6

The effect of simplification of a diagnostic prediction rule presented as a scoring system on performance

This is an accepted manuscript of an article published by Elsevier in the Journal of Pediatrics on the 13th of May 2015, available online: doi:10.1016/j.jclinepi.2015.05.006. © 2015.

Sanders S, Flaws D, Than M, Pickering JW, Doust J, Glasziou P. 2016. Simplification of a scoring system maintained overall accuracy but decreased the proportion classified as low risk. *Journal of Clinical Epidemiology*. 69:32-9, Elsevier, 2015.

This manuscript version is made available under the CC-BY-NC-ND 4.0 licence

<http://creativecommons.org/licenses/by-nc-nd/4.0/>(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

6.1 Preface to Chapter 6

The adoption of diagnostic prediction rules into practice is influenced by multiple contextual factors including potential users' perceptions about attributes of the tool itself. A clinical prediction rule may not be used if it is considered too much effort to apply or is perceived as being too complex or complicated in terms of both the number of predictors included, and the calculations required to derive an output. In the final study of this thesis, a diagnostic prediction rule is simplified using various methods. The comparative performance of the original and simplified prediction rules is evaluated using traditional performance measures and utility measures which offer insight into the clinical consequences arising from application of the alternate rules.

6.2 Abstract

Objective: Scoring systems are developed to assist clinicians in making a diagnosis. However, their uptake is often limited because they are cumbersome to use, requiring information on many predictors or complicated calculations. We examined whether, and how, simplifications affected the performance of a validated score for identifying adults with chest pain in an emergency department who have low risk of major adverse cardiac events.

Study design and setting: We simplified the Emergency Department Assessment of Chest Pain score (EDACS) by three methods: 1) giving equal weight to each predictor included in the score, 2) reducing the number of predictors and 3) using both methods – giving equal weight to a reduced number of predictors. The diagnostic accuracy of the simplified scores was compared with the original score in the derivation (n=1974) and validation (n=909) datasets.

Results: There was no difference in the overall accuracy of the simplified versions of the score compared with the original EDACS as measured by the area under the ROC curve (0.74 to 0.75 for simplified versions versus 0.75 for the original score in the validation cohort). With score cut-offs set to maintain the sensitivity of the combination of score and tests (ECG and c-troponin) at a level acceptable to clinicians (99%), simplification reduced the proportion of patients classified as low risk from 50% with the original score to between 22 and 42%.

Conclusion: Simplification of a clinical score resulted in similar overall accuracy, but reduced the proportion classified as low risk and therefore eligible for early discharge compared to the original score. Whether the trade-off is acceptable, will depend on the context in which the score is to be used. Developers of clinical scores should consider simplification as a method to increase uptake, but further studies are needed to determine the best methods of deriving and evaluating simplified scores.

6.3 Introduction

Clinical prediction rules are tools designed to improve clinical decision making. (2) They are often presented as simplified scoring systems. In such systems, the predictors (items from the patient's history or examination) have a point value which is summed to give an overall integer score for a particular individual. The scoring system may provide users with an estimate of the predicted risk of the outcome of interest for each of the integer scores (289, 290) and or may stratify an individual into a risk group (e.g. low, medium or high risk). A course of action may be implied or recommended based on this stratification (e.g. suitable or not suitable for early discharge). (291)

Despite increasing interest in the potential of scoring systems to augment the judgment of clinicians, the majority are rarely used. (105) One barrier to their use may be that the score is difficult to implement in practice. For example, the score may be cumbersome to use, requiring the collection of many pieces of information or information not normally collected during the consultation. (292) Further, it may be computationally complex with predictors of different point value, (293) of positive or negative value, or of non-integer point value. (294) Such characteristics also make the score difficult to remember and increase the potential for summing mistakes. Though presenting scoring systems as web based calculators or embedding them in electronic patient records may minimise these barriers, clinicians may be dissuaded from applying tools they believe are too complicated, not transparent or too much effort to apply.

Intuitively, scoring systems that contain only a small number of predictors and require only simple calculations to derive a result are more likely to be used and to be used correctly in practice. (295) It may be possible to simplify scoring systems to facilitate use in practice without loss of predictive power. For instance, a simplified version of the well-known Wells rule for acute pulmonary embolism has been shown in prospective validation studies to have similar diagnostic performance to the original rule. (296)

A scoring system to identify chest pain patients with a low risk of major adverse cardiac event (MACE) who could be discharged early from the emergency department (ED) has recently been developed and validated. (297) When used in conjunction with ECG and cardiac troponin tests as part of an accelerated diagnostic pathway, the combination of score and tests safely identifies just over half (51%) of the chest pain patients as low short term risk of MACE and suitable for discharge to early outpatient follow-up. This is a substantial improvement over existing accelerated diagnostic pathways which, at best, identify only 15% of chest pain

patients as safe for early discharge. (298) The scoring system, known as the Emergency Department Assessment of Chest-pain score (EDACS) includes 7 predictors. Each predictor has a different point value (Table 6.1) and two predictors have a negative value, that is, their value is subtracted from the score total. The variation in point values may make derivation of an individual's score somewhat onerous and subject to miscalculation.

We sought to simplify EDACS by 1) assigning equal weights to predictors included in the score, 2) reducing the number of predictors and 3) using both methods. We evaluated whether simplified versions of the score retained the original score's diagnostic accuracy and clinical usefulness in allowing early discharge of patients with chest pain from the emergency department whilst maintaining an acceptable level of sensitivity.

6.4 Methods

6.4.1 Derivation and validation of the original EDAC score

The EDAC score was derived in a population of 1974 consecutive patients aged 18 years and older who had had at least 5 minutes of symptoms consistent with acute coronary syndromes attending emergency departments in Brisbane, Australia and Christchurch, New Zealand between June 2007 and February 2010. Data were collected on prospective candidate variables and participants were followed for 30 days to determine the presence of MACE, defined as ST-elevation or non-ST-elevation myocardial infarction, emergency revascularisation procedure required, death from cardiovascular causes, ventricular arrhythmia, cardiac arrest, cardiogenic shock or a high atrioventricular block. Backward stepwise logistic regression was performed to identify a model which was converted into a scoring system by multiplying and rounding the model coefficients. The score originally included 6 predictors (male sex, age, diaphoresis, pain radiating to arm or shoulder, pain occurred or worsened with inspiration and pain reproduced by palpation). Based on clinician feedback on the acceptability of the score, the score was modified to include an additional predictor – presence of traditional cardiac risk factors or history of coronary artery disease (Table 6.1). Incorporation of this predictor meant that younger patients (≤ 50 years of age) would be assigned points if a history of coronary artery disease or 3 or more traditional risk factors (family history of premature CAD, dyslipidaemia, diabetes, hypertension or current smoker) were present.

In clinical practice, the score is used in conjunction with ECG and cardiac troponin tests performed at 0 and 2 hours (after presentation to the ED) as part of an accelerated diagnostic pathway (EDACS ADP). Based on a survey of clinicians who indicated that a 1 in 100 error rate

in discharging patients from the ED who have a MACE within 30 days due to unrecognised ACS was acceptable, (299) a cut-off of 16 was the optimal score to maintain sensitivity of the EDACS accelerated diagnostic pathway near 99% while maximising specificity. The EDACS ADP recommends discharge with follow-up investigation when the EDACS is <16 , there is no new ischemia on ECG and 0 and 2 hour troponin are both negative. Otherwise, patients are considered to be moderate to high risk and if no diagnosis is made, observed and a delayed troponin test conducted. The EDACS ADP was prospectively validated in a separate cohort of 608 patients attending the same emergency departments between October 2010 and December 2011. (297)

6.4.2 Development of simplified versions of the EDACS score

We developed simplified versions of the original EDACS (Table 6.1) using the derivation dataset described above and tested them in an extended validation dataset ($n=909$). The methods for simplifying the scores were:

1. Unweighted score: assigns either +1 or -1 point to the predictors with 1 point for each decade of life from 40 years.
2. Reduced scores: we limited the reduced scores to a maximum of 4 predictors based on research that suggested 3 or 4 pieces of information are all that can be held in working memory at one time. (300) To develop the reduced scores, beta coefficients from the logistic regression model were standardised and the 3 strongest predictors identified being age, gender and pain radiating to the arm. To maintain face validity, the predictor 'presence of traditional risk factors or history of coronary artery disease' was incorporated into the reduced scores.
 - (a) 'Reduced and weighted' score: the logistic model including the 3 strongest predictors was rerun and the beta coefficients were multiplied, rounded and doubled (to avoid half scores) to give different point values to each predictor.
 - (b) 'Reduced and unweighted' score: the predictors were assigned a value of +1, with age assigned 1 point for each decade of life from 40 years.

Scores for all patients in the derivation and validation data sets were calculated according to the original EDACS, 'Unweighted score', 'Reduced and weighted score' and 'Reduced and unweighted score'.

Table 6.1. Original EDACS and simplified scores

Predictors included in the score	Original EDACS		Simplified scores					
			Unweighted score		Reduced Scores			
					Weighted		Unweighted	
Age	Age	Points	Age	Points	Age	Points	Age	Points
	18-45	+2	40-49,	+1	18-45	+2	40-49,	+1
	46-50	+4	50-59,	+2	46-50	+4	50-59,	+2
	51-55	+6	60-69	+3	51-55	+6	60-69	+3
	56-60	+8	70-79	+4	56-60	+8	70-79	+4
	61-65	+10	80-89,	+5	61-65	+10	80-89,	+5
	66-70	+12	90-99	+6	66-70	+12	90-99	+6
	71-75	+14			71-75	+14		
	76-80	+16			76-80	+16		
	81-85	+18			81-85	+18		
	86+	+20			86+	+20		
Male sex		+6		+1		+9		+1
Age 18-50 and either; known coronary artery disease (previous acute myocardial infarction, coronary artery bypass graft or percutaneous intervention); or ≥ 3 Risk factors (smoking, hypertension, diabetes mellitus, hypercholesterolemia, family history coronary heart disease)		+4		+1		+4		+1
Diaphoresis		+3		+1				
Pain radiates to arm or shoulder		+5		+1		+8		+1
Pain occurred or worsened by inspiration		-4		-1				
Pain reproduced by palpation		-6		-1				
Total Score range		-8 to 34		-2 to 9		2-37		0 to 8

6.4.3 Evaluation of original and simplified versions of the EDAC score

We calculated the area under the Receiver Operating Characteristic (AUROC) curve for each version of the score in the derivation and validation datasets. We used the bivariate binomial method with adjustments for multiple comparisons to estimate the respective ROC curves and tested the differences between the areas for significance using STATA13. Calibration of the original and reduced models was evaluated in the validation dataset by dividing participants into risk deciles using the predicted probability for having MACE. In each of the deciles, the number of expected MACE cases (predicted) for both models was compared to the actual number of MACE cases (observed).

For each of the simplified versions of the score, we determined score cut-offs that maintained the sensitivity of the combination of score and tests at approximately 99% or greater. Using these cut-offs, we calculated the sensitivity, specificity and proportion of patients classified as low risk by each score when used in combination with tests (ECG and c-troponin). We considered a score clinically useful if it classifies >40% of patients with chest pain as low risk (which is higher than in previously reported studies).

We evaluated the ability of the alternative scores to classify patients to more appropriate risk categories. We calculated the event net reclassification improvement (NR_{Ievents}) as the net proportion of patients with events reassigned to a higher risk category by the simplified score in combination with tests, and the non-event net reclassification improvement (NR_{Inonevents}) as the net proportion of patients without events reassigned to a lower risk category. (301) We also calculated the net reclassification improvement (NRI) which reflects improved reclassification and is the sum of NR_{Ievents} plus NR_{Inonevents}. To account for differences in the relative importance of true positive and false positive classifications we calculate a weighted NRI. (302) For patients with suspected cardiac chest pain, a threshold of 2% (derived using the method of Pauker and Kassirer (53)), has been reported as the point of probability at which the risks from false positive testing are balanced with the risk of harm from untreated disease. (79) Based on this threshold, we determined the relative weight of true positives to true negatives to be 0.98 to 0.02 or 49:1 and used this weighting in the calculation of the weighted NRI. We used a bootstrap estimate of the variance of the statistic to calculate 95% confidence intervals for the NRI for events and non-events and the weighted NRI. (303)

6.5 Results

Age and sex distributions in the derivation and validation cohorts were comparable, as was the occurrence of MACE (Table 6.2). The presence of known coronary artery disease or 3 or more risk factors was slightly more common in the validation cohort as was the presence of the predictor 'pain on palpation'. Cardiac troponin tests were positive in 20% of the derivation cohort and 16% of the validation cohort.

In both the derivation and validation datasets, the AUROCs of the simplified scoring systems were similar to each other and to the original score (0.75 for the original score, 0.74 for the unweighted score, 0.75 for the reduced weighted score and 0.74 for the reduced unweighted score in the validation dataset; $p \geq 0.05$ for the difference between the original and simplified scores and for the difference between simplified scores) (Table 6.3). The predictions of both

the original and reduced models track the actual observed cases of MACE well in the validation dataset (Appendix C).

Table 6.2. Characteristics of the derivation and validation cohorts

Characteristic	Derivation cohort (n=1974)*	Validation cohort (n=909)
Mean age (range)	60.5 (19-98)	60.1 (21-95)
Male	1184 (60.0)	538 (59.2)
Known coronary artery disease or ≥3 risk factors	864 (43.8)	435 (47.9)
Age 18-50 and either known coronary artery disease or ≥3 risk factors	142 (7.2)	53 (5.8)
Diaphoresis	998 (50.6)	450 (49.5)
Pain radiates to arm	674 (34.1)	332 (36.5)
Pain on inspiration	427 (21.6)	216 (23.8)
Pain on palpation	166 (8.4)	114 (12.5)
Troponin positive if >0.04 ng/mL in QLD cases and >0.03 ng/mL in NZ cases at 0 or 2 hours)	393 (19.9)	145 (16.0)
ECG positive at 0 or 2 hours	135 (6.8)	76 (8.4)
MACE	305 (15.5)	133 (14.6)

QLD Queensland; NZ New Zealand; ECG electrocardiogram

*values are numbers (percentages) unless otherwise stated

At score cut-offs which maintained the sensitivity of the combination of the EDACS score, ECG and troponin tests at approximately 99%, simplification of the score decreased specificity in the derivation and validation cohorts (Table 6.3). In the validation cohort, the specificity of the original EDACS in combination with ECG and c-troponin tests was 58.1%. This was reduced to 47.4% for the unweighted score (cut-off <4, ≥4), 48.6% for the reduced weighted score (cut-off <17, ≥17) and 26% for the reduced unweighted score (cut-off <3, ≥3).

In the validation cohort, the original EDACS in combination with ECG and c-troponin classified 50% of patients as low risk (Table 6.3). This was reduced to 41% for the unweighted score (cut-off <4, ≥4), 42% for the reduced weighted score (cut-off <17, ≥17) and 22% for the reduced unweighted score. The 22% low risk classification by the reduced unweighted score is below the 40% minimum we pre-specified for clinical usefulness of a score. MACE occurred in 2 patients classified as low risk by the original and unweighted scores (cut-off <4, ≥4), in 1 patient classified as low risk by the reduced weighted score (cut-off <17, ≥17) and in no patients classified as low risk by the reduced unweighted score (cut-off <3, ≥3) (Table 6.4 and Figure 6.1). Among those classified as higher risk by the original score, 29% of patients had MACE. This was reduced to 24%, 25% and 19% in the unweighted, reduced weighted and reduced unweighted scores respectively (Figure 6.1).

Use of the simplified scores resulted in more patients being incorrectly reclassified than correctly reclassified. Use of the unweighted score did not result in the reclassification of any

individuals with MACE to the higher risk category (NRlevent 0%) but a net 10.7% of individuals without MACE were incorrectly reclassified to the higher risk category. For the reduced weighted score, 1 patient with MACE was reclassified to the higher risk category (NRlevent 0.75%) but a net 9.55% of individuals without MACE were incorrectly reclassified to the higher risk category. With the reduced unweighted score 2 individuals with MACE were correctly reclassified to the higher risk category (NRlevent 1.5%), however a net 31.8% of individuals without MACE were also incorrectly reclassified to the higher risk category. Even when taking into consideration the relative importance of failing to detect a MACE in a patient classified as low risk and classifying a patient who does not have MACE as high risk (failure to detect MACE was considered 49 times more important), the simplified scores do not improve classification (wNRI -2.8 and -16.2 for the reduced weighted and reduced unweighted scores) (Table 6.5 and Appendix C).

Table 6.3. Area under the receiver operator characteristic curve (AUROC) for the original and simplified versions of the score and sensitivities and specificities when the score is used in conjunction with ECG and c-troponin tests

	Original score			Simplified scores			
	Unweighted score			Reduced weighted score		Reduced unweighted score	
	Cut-off ≥16	Cut-off ≥3	Cut-off ≥4	Cut-off ≥16	Cut-off ≥17	Cut-off ≥3	Cut-off ≥4
Derivation dataset (n=1974)							
Sensitivity	99.0	100	99.0	99.3	98.7	100	99.7
(95% CI)	(97.2-99.7)	(98.8-100)	(97.2-99.8)	(97.7-99.9)	(96.7-99.6)	(98.8-100)	(98.2-100)
Specificity	49.9	22.4	40.5	38.0	42.4	24.0	46.4
(95% CI)	(47.5-52.3)	(20.4-24.4)	(38.1-42.9)	(35.7-40.4)	(40.0-44.8)	(22.0-26.2)	(44.0-48.8)
AUROC	0.74	0.73 (0.70-0.76)		0.72 (0.69-0.74)		0.71 (0.68-0.74)	
(95% CI)	(0.71-0.76)	p=0.20		p=0.02		p=0.02	
and p value*							
Validation dataset (N=909)							
Sensitivity	98.5	99.3	98.5	99.3	99.3	100	97.7
(95% CI)	(94.7-99.8)	(95.9-99.9)	(94.7-99.8)	(95.9-99.9)	(95.9-99.9)	(97.2-100)	(93.5-99.5)
Specificity	58.1	25.4	47.4	42.5	48.6	26.3	50.1
(95% CI)	(54.6-61.6)	(22.4-28.6)	(43.9-51.0)	(39.0-46.1)	(45.0-52.2)	(23.2-29.5)	(46.6-53.7)
AUROC	0.75	0.74 (0.69-0.78)		0.75 (0.71-0.79)		0.74 (0.70-0.78)	
(95% CI)	(0.70-0.79)	p=0.28		p=0.98		p=0.96	
and p value*							
Percent classified as low risk (95% CI)†	49.8 (0.47-0.53)	21.8 (0.19-0.25)	40.7 (0.38-0.44)	36.4 (0.33-0.40)	41.6 (0.38-0.45)	22.4 (0.20-0.25)	43.1 (0.40-0.46)

*p value is for the difference between the original and simplified versions of the score; †percent classified as low risk = FN+TN/total N

Table 6.4. Proportion of patients assigned to low and high risk categories by the scores when used in combination with ECG and c-troponin tests by event in the validation dataset (n=909)

Risk category	Original score		Simplified scores					
			Unweighted score (cut-off ≥ 4)		Reduced weighted score (cut-off ≥ 17)		Reduced unweighted score (cut-off ≥ 3)	
	Non event	Event	Non event	Event	Non event	Event	Non event	Event
Low risk	451 (58%)	2 (1.5%)	368 (47%)	2 (1.5%)	377 (49%)	1 (0.75%)	204 (26.3%)	0 (0%)
Higher risk	325 (42%)	131 (98.5%)	408 (53%)	131 (98.5%)	399 (51%)	132 (99.25%)	572 (73.7%)	133 (100%)
Total	776 (100%)	133 (100%)	776 (100%)	133 (100%)	776 (100%)	133 (100%)	776 (100%)	133 (100%)

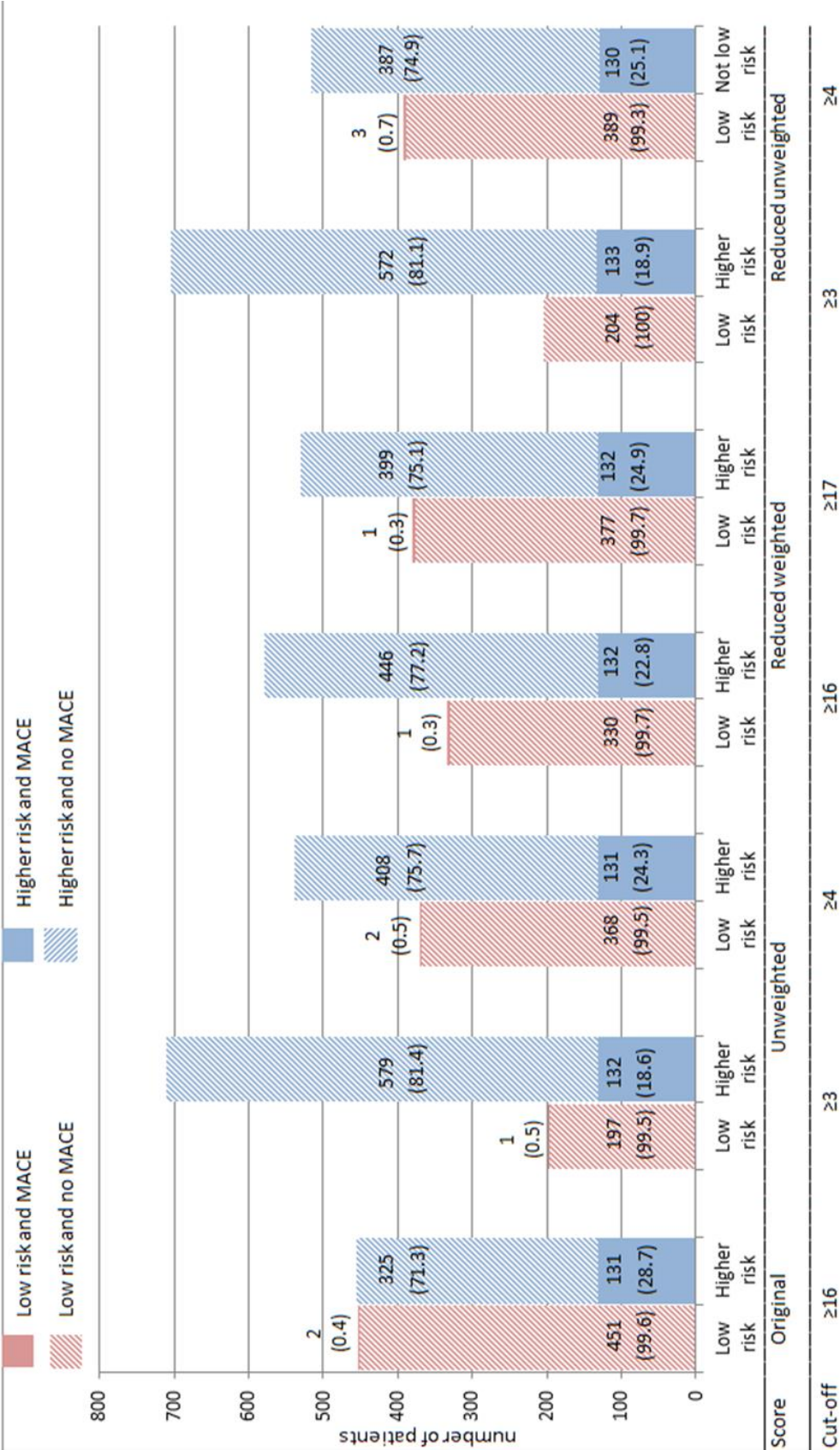
Table 6.5. Changes in classification with simplified scores when used in conjunction with ECG and c-troponin tests

	Unweighted Score (cut-off ≥ 4)	Reduced weighted Score (cut-off ≥ 17)	Reduced unweighted score (cut-off ≥ 3)
Event NRI net % (95% CI)	0	0.75 (0 to 3.9)	1.5 (-0.6 to 3.6)
Non-event NRI net % (95% CI)	-10.7 (-13.1 to -8.51)	-9.5 (-12.4 to -7.1)	-31.8 (-37.2 to -28.4)
NRI*	-10.7	-8.8	-30.3
weighted NRI (95% CI)	NA‡	-2.8 (-13.8 to 8.2)	-16.2 (-27.9 to -6.3)
Event to non event merit 49:1†			

NRI net reclassification improvement

* a negative value indicates worse classification; † correctly reclassifying patients with an event is considered 49 times more important than reclassifying patients without events; ‡ a weighted NRI was not calculated as there were no event reclassifications with this score

Figure 6.1. Comparison of original and simplified scores used in combination with ECG and c-troponin tests in validation dataset (n=909)



6.6 Discussion

This study has found that simplifying EDACS to facilitate use in practice comes at the cost of decreased specificity. Though there was no difference between the original and simplified scores in overall accuracy in the validation dataset, at cut-offs set to maintain sensitivity at around 99%, all three simplified scores when used in combination with ECG and c-troponin tests, were less specific than the original score with the proportion of patients who could be discharged early reduced from 50% to 43% at best. For two of the simplified scores (the reduced weighted and reduced unweighted score), the proportion classified as low risk was above the pre-specified clinically relevant minimum, and higher than the proportion of patients classified as low risk by existing scores. (291, 304)

Whether the trade-off between simplicity and performance is acceptable, requires consideration of the consequences of both and the context within which the score is to be used. Application of the recommendations from the simplified score instead of the original EDACS would mean that more patients would undergo further investigation and observation. Prolonged observation and investigation of patients with chest pain contributes to emergency department overcrowding which is associated with high costs and adverse outcomes including increased mortality. (305) However, this should be balanced against the possible benefits of increased application of the accelerated diagnostic pathway which may occur with the simplified scores as they are easier to apply (information on fewer predictors is required) or calculate. (306) In the context of the assessment and disposition of chest pain patients, increased use of the EDACS accelerated diagnostic pathway could have major benefits. Recently, a randomised controlled trial found that a similar accelerated diagnostic pathway, in which use and application of the results of the pathway were at the discretion of the attending clinician, almost doubled the proportion of patients with chest pain who were successfully discharged within 6 hours of presentation at the emergency department. (201) Thus, even a small increase in the number of clinicians using the accelerated diagnostic pathway and then applying its recommendations could have significant benefits in terms of reduced consumption of health resources, costs and patient inconvenience.

Simplification of the scoring system may also lead to more accurate implementation. Currently EDACS is provided to clinicians as a paper based score where clinicians manually compute the result. Given the need to add and or subtract the different point value of several predictors, some calculation error is inevitable. Simplifications that minimise the computational effort required to derive the result may result in fewer calculation errors and more accurate implementation of EDACS in clinical practice.

Although it could reasonably be expected that incorporating a less burdensome score would facilitate use, it could potentially have the opposite effect by reducing the face validity of the score. Scores with a reduced number of predictors may not contain predictor variables clinicians believe are important, reducing clinicians trust in it. Although we incorporated the predictor variable 'traditional risk factors or history of coronary artery disease' which was not statistically significant in the multivariable model but was considered by clinicians when absent from the score to decrease its credibility, we did exclude three other predictors from the reduced score (diaphoresis, pain occurred or worsened with inspiration and pain reproduced by palpation) that are commonly used in clinical practice, but which were not statistically significant predictors in the multivariable analysis. Attitudes that scoring systems over simplify the clinical assessment process, disrespect clinical complexity and the belief that more 'information' (predictor variables) will improve prediction, may be a barrier to use of scoring systems in general and simplified scoring systems in particular. To determine the acceptability of the simplified versions of the score to clinicians, and their willingness or unwillingness to trade off accuracy for simplicity requires further study.

We investigated 3 methods of simplifying the EDAC score: 1) assigning equal (unit) weights to each predictor included in the score, 2) reducing the number of predictors and 3) using both methods. At score cut-offs that maintain sensitivity at approximately 99%, equally weighting the predictors or reducing the number of predictors but maintaining different predictor weights reduced specificity to a similar degree. When both methods were applied (reducing the number of predictors and equally weighting these predictors) at a score cut-off that maintained adequate sensitivity, score specificity was reduced to a level below the 40% minimum specificity we considered clinically useful.

Equal (unit) weighting is a common method of simplifying scoring systems and has been shown in a variety of contexts inside and outside healthcare to produce models yielding predictions that correlate highly with models using so called optimal weights (predictors 'weighted' to optimize the relationship between the prediction and outcome). (307) It has also been suggested that equally weighted models are more transferable to other settings because their weights are not specific to the population in which they were derived. (36) This method of simplification has been applied to the well-known Wells score for pulmonary embolism. In contrast to the findings of this study, simplification of the Wells PE score did not reduce the proportion of patients with PE who had been safely classified as low risk compared to the original score. (308) For the EDAC score, giving equal value to each of the predictors reduced

the range of possible scores substantially from a range of -8 to +34 for the original score to a range of -2 to 9 for the unweighted score. This lack of graduation between potential score cut points may explain in part the reduction in specificity of the simplified scores and is a possible limitation of this method of simplification.

Reducing the number of predictor variables in the score is another method of simplifying scores to facilitate their use in practice. Such scores require less information and effort to derive a result. Typically in prediction models, a few variables with strong effects account for most of the predictive power, with the remaining weaker variables contributing relatively little. (309) While some loss of predictive value is inevitable when the relatively 'weaker' predictors are removed, such scores may still perform well in validation data because the potential for over fitting is reduced. In this study, the simplified scores with a reduced number of variables performed reasonably well in the validation data, but only when weighting of the predictors was maintained.

The area under the receiver-operating characteristic curve is a familiar and commonly used measure to compare prediction models. However, it has notable limitations, including insensitivity to changes in model performance. (310) Further, it focuses solely on accuracy without incorporating information on consequences. While we found no difference according to the AUROC between the original and simplified scores, from a clinical perspective, this analysis provides little useful information to inform model selection. In practice, the main aim of the EDACs score in patients with chest pain is to distinguish individuals in whom MACE can safely be excluded (estimated probability of MACE <2%) from those who should undergo further observation and testing. As such, we focus on the more clinically useful paired summary statistics (sensitivity and specificity) which express the magnitude of false negative and false positive test errors, and proportion classified as low risk with original and simplified models. Further, we assess the ability of the original and simplified models to more appropriately reclassify individuals into risk categories taking into account the relative value of false negative and false positive reclassifications.

As yet, the simplified EDAC scores have only been temporally validated. A further limitation of our study is that the results reported are derived from the application of the scoring system to the data without the influence of the clinician. In practice, the EDACS score will be used in conjunction with clinical judgment. We do not know what effect this might have on the accuracy of the original or simplified scoring systems presented here. Ideally, further validation of the simplified scores would be conducted in different settings and by researchers

not involved in their development. (5) In addition, we suggest that future validations should also contrast the accuracy of the simplified scores with the current diagnostic pathway. This may be clinical judgment alone or a standard departmental protocol. If the scores are found to be more accurate than clinical judgment, further evaluation could be considered. Such studies would determine the extent to which the scores are actually used and whether this use leads to improved outcomes for patients. The performance of the original EDACs score when used as an adjunct to clinical judgment (as part of an accelerated diagnostic pathway) is currently being evaluated in this manner (Australian New Zealand Clinical Trials Registry Number 12613000745741). An interim step may be a modelling study to determine the incremental value of the scores over clinical judgment alone, similar to that conducted by Broekhuizen et al. (12)

The complexity of a clinical prediction rule can be a barrier to uptake and simplification may assist. As this case illustrates, there may be a loss in the performance characteristics of the score, but in some contexts, a small loss may be acceptable if uptake is increased. Our study describes several means by which scores can be simplified and illustrates one approach to their evaluation. Many alternative and complementary methods of evaluation exist. A general methodology for performing and assessing the effects of simplification of prediction rules presented as scoring systems would be a useful supplement to the existing detailed methodological guidance on prediction tools development and evaluation.

6.7 Conclusion

Developers and evaluators of scoring systems should explore the effects of simplification on predictive performance. Where there is a trade-off between complexity and accuracy of the scoring system, the implications of each can be considered. The decision on which score to use (the original or simplified) will be context dependent; in some circumstances it may be acceptable to have some error if the score is more likely to be used and to be applied consistently and correctly.

6.7.1 Author contributions

Conceived and designed the study: SS, JD, PG. Analysed the data: SS, DF. Wrote the first draft of the manuscript: SS. Contributed to writing of the manuscript: SS, DF, MT, JWP, JD, PG. Agree with manuscript results and conclusions: SS, DF, MT, JWP, JD, PG. International Committee of Medical Journal Editors (ICMJE) criteria for authorship read and met: SS, DF, MT, JWP, JD, PG.

SS received funding from an Australian Postgraduate Award scholarship and the Screening and diagnostic Test Evaluation Program which is supported by a National Health and Medical

Research Council Program Grant (<https://www.nhmrc.gov.au/>). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Chapter 7 Discussion and Conclusions

7.1 Preface to Chapter 7

This final chapter presents a summary of the thesis findings, explores their limitations and strengths within the context of the wider literature, outlines implications and makes suggestions for future research.

7.2 Summary of findings

The first systematic review of this thesis aimed to determine the comparative performance of diagnostic prediction rules (when applied to study data independently of clinical judgment) and clinical judgment against a common reference standard. The review sought to address the limitations of existing studies comparing clinical prediction rules and clinical judgment by specifically reviewing diagnostic prediction rules intended for use in clinical practice and by reporting comparative performance using metrics against which the clinical value of the alternate methods can be judged.

This systematic review (reported in Chapter 2) found no clear evidence for the superiority of either clinical judgment or diagnostic prediction rules. In none of the comparisons did diagnostic prediction rules both classify a larger proportion of patients as not having disease (thereby avoiding further testing or treatment) and miss fewer cases of disease. In one comparison, clinical judgement both classified a larger proportion of patients as not having disease and missed fewer cases of disease. In all other comparisons there was a trade-off between the proportions classified as not having disease and the proportion of missed diagnoses. For example, diagnostic prediction rules for the diagnosis of deep vein thrombosis were generally more efficient and classified a larger proportion of all participants as not having disease (therefore avoiding further testing) than clinical judgment, but this was often at the expense of missed diagnosis of deep vein thrombosis as determined by the reference standard. The differences between the two methods of judgment most likely represent a different threshold for positivity, with the clinical value of either method dependent upon the relative benefits and harms of missed diagnosis versus avoidance of further testing or treatment.

The second systematic review of this thesis aimed to determine the effect of diagnostic prediction rules on patient and process outcomes. It was the first study to systematically review randomised trials comparing the effect of care provided with and without a diagnostic prediction rule across a range of clinical conditions. The review also examined the frequency of reporting of intervention characteristics and implementation methods necessary to interpret study findings and enable replication of study findings in practice.

This systematic review (presented in Chapter 3) of 25 randomised trials found that few studies reported patient outcomes as the primary study outcome, but that diagnostic prediction rules have a positive effect on process outcomes for some clinical conditions. Exposure to diagnostic prediction rules for Group A Streptococcal throat infection reduced symptoms in the one study

reporting this outcome and reduced antibiotic prescribing in a meta-analysis of five trials. Diagnostic strategies for cardiac chest pain incorporating a diagnostic prediction rule improved early discharge rates (one study) and decreased hospitalisations (one study). The effects of the Ottawa Ankle Rules on process outcomes were positive when used as an add-on test in conjunction with clinical judgment (one study) but not when it was used as a triage test prior to clinical examination (one study). Diagnostic rules for acute appendicitis reduced time to therapeutic operation (one study) and showed a non-statistically significant reduction in non-therapeutic operations in a meta-analysis of five studies. However, diagnostic prediction rules for children with fever did not improve the process outcomes measured. Details of the study interventions, particularly the control group interventions and the methods of implementing the diagnostic prediction rule, were infrequently reported.

The third study of this thesis (presented in Chapter 4) aimed to derive and externally validate a diagnostic prediction rule for differentiating children presenting to the primary care setting with serious bacterial infection from those with self-limiting infection using an existing dataset. The derived prediction rule was also to be used as the basis for further studies to determine the added value of the inflammatory biomarker C-reactive protein, and the effect of prediction rule simplification on performance. However, as described in Chapter 4, efforts to derive the prediction rule were not successful. Due to the volume and likely nature of the missing data in the sourced dataset and concerns regarding the methods of dealing with missing data that were available at the time, I determined that derivation of a clinically sensible and valid tool could not be achieved. Consequently, the study of the added value of C-reactive protein over features from the history and clinical examination could also not be performed, and the study of the effect of simplification required modifications to the research plan.

The fourth study, and third systematic review of this thesis (reported in Chapter 5), aimed to determine the diagnostic accuracy and independent value of the biomarker C-reactive protein in the primary care setting. At the time, C-reactive protein was widely used as a laboratory and point of care test in some countries and was being heavily promoted in others, yet its accuracy and independent value had not been established in the non-hospitalised paediatric population.

This systematic review found that C-reactive protein provided moderate and independent information for both ruling in and ruling out serious bacterial infection in infants and children presenting with fever. All included studies were based in emergency department settings. From six of seven studies included in the review, the pooled positive likelihood ratio of 3.64 (95%CI 2.99 to 4.43) represented a small increase in the probability of serious bacterial

infection with a positive C-reactive protein test, and the negative likelihood ratio of 0.29 (95%CI 0.22 to 0.40) a moderate decrease in the probability of a serious bacterial infection with a negative CRP test. From five multivariable modelling studies, C-reactive protein was an independent predictor of serious bacterial infection, that is, when included in a model with other covariates, it had a statistically significant association with the outcome of serious bacterial infection. While this review suggests that measuring C-reactive protein is a helpful step in the diagnostic workup of non-hospitalised children, the modest likelihood ratios confirm the importance of assessing the result of the test in the light of clinical findings.

The final study presented in this thesis (Chapter 6) aimed to determine the effect of different methods of simplifying a diagnostic prediction rule presented as a scoring system on performance. The complexity of clinical prediction rules has been identified as a barrier to the implementation of prediction rules in practice, but the effects of different methods of simplification on diagnostic accuracy and risk classification had not been explored. As discussed earlier in this chapter (page 140) and thesis (Chapter 4), this study was to be conducted using the CPR developed for identifying children with serious infection. When this was not possible, and an attempt to source other suitable data in the clinical area of interest was unsuccessful, data from a derivation and validation study of a prediction rule for the identification of chest pain patients with a low risk of major adverse cardiac event were utilised.

In this study, methods of simplifying the prediction rule maintained overall accuracy as measured by the area under the AUROC but reduced the proportion of patients classified as low risk and led to worse reclassification. Three methods of scoring system simplification were tested: (1) giving equal weight to each predictor included in the score; (2) reducing the number of predictors; and (3) giving equal weight to a reduced number of predictors. The accuracy of the simplified and original scores, as measured by the area under the ROC was similar. However, when the original and simplified scores were used as part of a diagnostic pathway with ECG and cardiac troponin tests, and thresholds set at a pre-specified minimum sensitivity of >99, all simplification methods reduced the proportion of patients classified as low risk of major adverse cardiac events and therefore eligible for early discharge compared to the original score. The proportion of patients classified as low risk by diagnostic pathways that included the scores simplified using method 1 (equal weighting) or 2 (reduced number of predictors) was above the 40% threshold we had pre-specified as being the minimum required to be clinically useful. However when the score was simplified by method 3 (equal weighting

and reduced number of predictors) and used in the diagnostic pathway, the proportion classified as low risk was below this threshold (22%). Use of the simplified scores also resulted in more individuals being incorrectly reclassified into risk groups than correctly reclassified. This was the case even after taking into consideration the relative importance of failing to detect a major adverse cardiac event in a patient classified as low risk, and classifying a patient who did not have an event as high risk (missing a diagnosis of a major cardiac event was considered more important than the possible harm from over testing or treatment in someone without an event).

7.3 Limitations and strengths of the studies and thesis in the context of the wider literature

There are a number of important strengths of the thesis research projects. The systematic reviews of the comparative performance of diagnostic prediction rules and clinical judgment and the effect of care provided with and without a diagnostic prediction rule are the first in this topic area. They address a gap in the existing literature regarding the potential for diagnostic prediction rules to assist clinicians' judgments, and to affect process and patient outcomes. Together these reviews give an overview of the current status of diagnostic prediction rules in the context of their stated potential as powerful tools to improve clinical decision making. Further, the reviews support calls for prediction research to shift from the development of more prediction rules, to rigorous evaluation of existing ones. (105, 311) By critiquing the design, conduct and reporting of the studies included in these reviews, areas for improvement in this field of research were also identified.

Key strengths of the systematic review of the accuracy and independent value of C-reactive protein lie in its novelty and timeliness. At the time this review was conducted and published (2007-2008), C-reactive protein was widely used as a laboratory test in tertiary settings and routinely as a point of care test in some countries, and was being promoted or considered for use by primary care clinicians in other countries to assist in differentiating between children with a bacterial infection that may benefit from the use of antibiotics, and those with a non-bacterial or self-limiting infection. However, its accuracy had not been evaluated using a systematic review methodology. Diagnostic tests that are introduced into practice without rigorous evaluation of their accuracy, a pivotal component in the evaluation of diagnostic tests, can lead to unwanted clinical consequences and increased healthcare costs arising from unnecessary testing. Using what was then emerging guidance for the conduct of systematic reviews of diagnostic test accuracy, but what is now recognised as the key guidance for this type of review, and employing a novel statistical method for estimating average sensitivity and

specificity, (312) this review informed the use and further evaluation of this particular biomarker in the primary care setting.

While the study of the effects of prediction rule simplification on performance provides information on the effects of simplification of a prediction rule for a specific clinical indication, the strength of this study lies in its contribution to the existing literature related to the development, evaluation and implementation of diagnostic prediction rules. It extends the limited existing research on prediction rule simplification by examining different methods of simplification, by demonstrating one approach to the development and evaluation of simplified scores and by examination of the context-dependent clinical implications of simplification. A particular strength of this study was the use of reclassification and utility based metrics to assess the performance of the alternate models in addition to traditional performance measures and, in this way, provide more information on the clinical usefulness of the simplified rule.

There are important limitations to each of these research projects, as identified explicitly in the preceding chapters. The findings of the systematic reviews of diagnostic prediction rules are primarily limited by the number and nature of the included studies. The conclusions of these reviews are based on few studies that were often judged to be at unclear or high risk of bias which may lead to the systematic over or underestimation of the test accuracy or treatment effect. (313-316) Further, the design features of many of the studies included in both the reviews has implications for interpretation of the study results. In the review of the comparative performance of diagnostic prediction rules and clinical judgment, for instance, study participants' knowledge of the study design was likely to have influenced the probability estimates, diagnoses, intended or actual actions of clinicians upon which comparative performance is judged. For both reviews, screening of titles and abstracts was only performed by one reviewer, with only a small proportion checked by a second reviewer. Thus, it is possible that studies relevant to both reviews may have been overlooked despite implementation of other methods to minimise this possibility. The conclusions of the impact review are likely to be context specific and limited by inadequate sample size in many of the included studies. Furthermore, as judgments regarding the reporting of the study interventions and implementation methods in the impact review were based only on the presence of any description, rather than the adequacy of the description, the review is likely to have overestimated the reporting quality of the included studies. While the review of the comparative performance of diagnostic prediction rules and clinical judgment also aimed to

determine the added value of diagnostic prediction rules beyond that obtained from the clinicians' implicit judgment, by including studies comparing multivariable prediction models of clinical judgment with and without a prediction rule, no eligible studies with this objective were identified.

For the systematic review of the diagnostic accuracy and independent value of C-reactive protein, an important limitation arose from the absence of studies conducted in the primary care setting. All included studies were performed in the emergency department where the prevalence of serious bacterial infection is usually higher. In the lower prevalence primary care setting, C-reactive protein may be less useful for changing the pre and post-test probability of serious infection, as sensitivity and specificity have been shown to vary by prevalence, with lower test accuracy often seen in populations where the prevalence of the target condition is low. (78)

The principle limitation of the study assessing the effects of diagnostic prediction rule simplification on performance was that the simplified scores have been only temporally validated. The more the validation cohort differs from the derivation cohort, the stronger the test of generalisability of the prediction rule. (62) While the populations in which the simplified rules were derived and validated were separated by time, the cohorts remained similar, sharing the same inclusion and exclusion criteria and the same predictor and outcome definitions and measurement methods. A further limitation arose from the derivation and validation of the simplified rules being performed by the same author. While evaluating one's own prediction rule is a useful first step, it is less desirable than an independent evaluation conducted by researchers not involved in their derivation. (105, 317) Lastly, the act of simplification is predicated on the assumption that simplified scores are more likely to be adopted and used completely and correctly in practice. While some research suggests that complexity and useability are barriers to the use of some clinical prediction rules (95, 96) and other clinical innovations such as clinical practice guidelines (318) and clinical decision support systems, (319) as far as I have been able to determine there is no research evaluating the relationship between prediction rule complexity and actual use.

The limitations of this thesis in addressing the aims of the thesis as a whole derive primarily from my decision not to continue with the derivation of a diagnostic prediction rule for the identification of serious bacterial infection in children. As a consequence, the second goal of the thesis (to assist primary care clinicians' management of children with possible serious infection) could only be partially achieved, and modification of some aspects of the proposed

research plan was necessary. My decision not to proceed with the derivation was a pragmatic one, reached after preliminary analysis of the sourced dataset revealed that, whilst a prediction rule *could* have been derived, for reasons previously described (Chapter 4), I judged a valid and clinically sensible rule could not. The derivation of prediction rules for the sake of publication, that are suboptimal in terms of their validity and clinical sensibility, have been justifiably criticised. (102) The existence of such prediction rules in the literature conceivably undermines clinicians' confidence in prediction rules as a whole and, further, may adversely affect implementation of potentially useful clinical prediction rules. Ultimately, they are unlikely to be accepted or implemented by the very clinicians that they are intended to assist.

7.4 Implications and recommendations arising from this thesis

The review of the literature related to the development and evaluation of diagnostic prediction rules over the duration of this thesis suggests that the literature, while dramatically increasing in volume, is failing to produce a commensurate amount of useful knowledge about the clinical value of these tools. The key implications arising from the findings of this thesis, general recommendations for research practice and one recommendation specific to childhood infections are discussed in detail below.

General recommendations for diagnostic prediction rule research practice:

- further investigation of the comparative performance of diagnostic prediction rules and contemporary diagnostic strategies in both the early and later stages of diagnostic prediction rule evaluation
- improved design, conduct and reporting of comparative studies of diagnostic prediction rules
- investigation of methods of developing and evaluating simplified prediction rules and investigation of whether, and how, the complexity and the face validity of a clinical prediction rule affects uptake

Recommendation specific to the investigation of children with possible serious bacterial infection:

- further investigation of the value of the biomarker C-reactive protein for identifying children with serious bacterial infection in the primary care setting

7.4.1 Further investigations of the comparative performance of diagnostic prediction rules and contemporary diagnostic strategies in both the early and later stages of diagnostic prediction rule evaluation

Despite calls for research directly evaluating, in the same patients or study population, alternate diagnostic tests that might be used at the same place in the diagnostic pathway (124, 320), the reviews of diagnostic prediction rules presented in this thesis indicate that such analyses are infrequently conducted in the field of prediction rule research. Evaluations of the added value of diagnostic prediction rules are also rare. Elucidating the comparative diagnostic performance of diagnostic prediction rules and clinical judgment and/or the added value of diagnostic prediction rules over clinical judgment is an important step in the early phase evaluation of diagnostic prediction rules, both for informing of the role of the prediction rule and guiding further evaluation, and for facilitating the acceptance and amenability of clinicians to further evaluation efforts. In the absence of direct comparative research, judgments must be made on the basis of indirect comparisons that are prone to confounding effects due to differences in patient groups and study methods. (321) However, evaluation of comparative performance adds complexity to the design and conduct of validation studies and practical guidance may be necessary to ensure 'fair' comparison (for example, the prediction rule and clinical judgment are compared at a similar time point in the diagnostic pathway). In the later stages of evaluation, comparative studies of care, provided with and without a prediction rule, are imperative for assessing the effect of prediction rules on health outcomes. While the existing clinical prediction rule development framework places comparative studies at the centre of later stage evaluations of clinical prediction rules, with randomised comparisons being the optimal design, the review presented in this thesis and other reviews of prediction rules in specific clinical areas (100, 106, 123), indicate that few studies of this type are being conducted. This is likely to be a reflection of the challenging nature of this kind of evaluation, but may also suggest a lack of appreciation or understanding among developers and or potential users, of the steps required between the development of a prediction rule and its use in practice. Despite their limitations, modelling studies may be a useful alternative, or an interim step, to the conduct of randomised impact studies. (93, 322)

7.4.2 Improved design, conduct and reporting of comparative studies of diagnostic prediction rules

The systematic reviews of diagnostic prediction rules presented in this thesis identified numerous challenges to the design and conduct of comparative studies when they are undertaken. Many of the threats to the validity of these studies may be minimised or ameliorated by researchers following existing guidelines for the conduct and reporting of

diagnostic accuracy studies (STARD) and randomised trials (CONSORT), including extensions of these guidelines for studies of nonpharmacologic treatments and cluster trials. (178, 209) Though the existence of these guidelines or their endorsement by journals has led to improvement in reporting quality (323, 324), this improvement is modest and further enforcement of adherence to reporting guidelines among researchers, editors and peer reviewers may be necessary to improve the quality of reporting of comparative studies of prediction rules to a satisfactory level. Additional design considerations for studies comparing prediction rules to clinical judgment or usual care may be needed to minimise the risk of study-induced behavioural change.

The review of the comparative performance of diagnostic prediction rules and clinical judgment has identified a number of methodological limitations and nuances within the included studies that impact on the observable comparative performance of the two approaches that should be further considered in future work in this area. Studies should include a more detailed discussion of the implications of the context of the comparison and factors that affect the location of test and treatment thresholds. For example, a CPR with a high sensitivity threshold may be being compared to clinical judgment in a context where ruling out the disease may not be the only consideration. Factors such as the seriousness of the target condition, the treatment options available, cost availability and side effects of diagnostic tests should also be discussed as factors that affect the test-treatment thresholds and consequently the interpretation of comparisons between CPRs and clinical judgment. (52) As it is not known which study characteristics are associated with diagnostic performance of prediction rules, further exploration of the heterogeneity in studies of comparative diagnostic performance should be undertaken to generate hypotheses for further testing. This may include analysis of subgroups according to prevalence, clinicians' experience, thresholds for ruling in and ruling out, and clinical conditions. Within clinical conditions, the effect of different prediction rules could be explored. The size of the studies and their ability to detect a clinically important difference between diagnostic prediction rules and clinical judgment to assist users of the review in interpreting the individual study and review findings should be assessed. In this review the outcome measures of 'safety' (false negative rate) and 'effectiveness' (proportion of study participants classified as not having disease) have been used. These are less conventional measures than the standard specificity and sensitivity, but were used because of their greater applicability in the clinical context. Further work is needed on methods of reporting the study finding and means for facilitating users understanding of the primary review outcomes. Pooling of studies according to the difference (in the proportion of

missed cases of disease among those classified as not having disease) between the two alternatives, or using metrics summarising the difference (such as the median) are possible alternatives or additions.

The review of the impact of diagnostic prediction rules has revealed a clear need for trials in this area firstly to clearly define how diagnostic prediction rules might be expected to affect patient and clinical process outcomes, and then to demonstrate these effects via the measurement and reporting of: 1) patient centred outcomes that ensure the effects of diagnostic prediction rules on all study participants are assessed; and 2) clinical process outcomes that allow the resulting health effects to be interpreted and translated into practice. Identifying the many mechanisms by which a diagnostic prediction rule may alter patient health, and the complex interactions between them, is likely to be a difficult task. It is, however, necessary to facilitate a full evaluation of the diagnostic strategy. A practical tool may assist in formulating a clear scientific rationale for the intended effects of diagnostic prediction rules by comparison with the existing diagnostic pathway. (50) The use of such a framework may highlight the need for studies to measure more patient and process outcomes than they currently do, which is likely to be both costly and challenging.

The review of the impact of diagnostic prediction rules also found that reporting of intervention characteristics is poor, and needs to be improved if results are to be used to enhance diagnostic practice. Comparator strategies described as 'standard care' or 'conventional investigation' were particularly poorly reported. Though such strategies may be difficult to translate into a prescriptive format, they are equally under evaluation in studies of the effect of diagnostic prediction rules, and not reporting them poses an irrevocable impediment to the interpretability of results. Furthermore, elements of implementation were rarely reported in these studies, making it difficult to determine the mediating effects of implementation factors on the study outcomes. Reporting such information is likely to be difficult under the space constraints of trial publication, but supplementary reports or graphical presentations may be an effective way to communicate these important details. (325)

There are many questions remaining about the effect of diagnostic CPRs. Firstly, it is unclear which factors differentiate between diagnostic prediction rules that improve patient or process outcomes and those that do not. Characteristics of the prediction rules themselves, for example, whether they are directive or assistive, or the way they are delivered could be further explored, though the feasibility of such analysis will be influenced by the number of

available studies. Secondly, just as medical interventions (pharmaceutical interventions or surgical procedures) when used as intended may cause harm, it is possible that the proper implementation of effective prediction rules may also contribute to undesirable outcomes (such as decreased clinician satisfaction, clinical reasoning ability or patient satisfaction with their clinician). Such potential harms should be further explored.

7.4.3 Investigation of methods of developing and evaluating simplified prediction rules and investigation of whether, and how, the complexity and the face validity of a clinical prediction rule affects rule uptake

The complexity of clinical prediction rules may be a modifiable barrier to their adoption, and developers and evaluators of prediction rules should derive, test and consider the clinical implications of simplified versions. As the simplification study presented in this thesis suggests, simplification may affect performance, but examination of the trade-off between usability and performance may have different implications in different situations. Given the fundamental role that clinicians play in the implementation of prediction rules, developing a prediction rule should not simply be a matter of identifying and publishing the most finely tuned model, but should be a clinician-centred process that considers the practical aspects of applying the tool in clinical practice. While assessing the effect of simplification of a prediction rule on performance can be done within typical validation and derivation studies and does not add significantly to the complexity of the evaluation process, obtaining an understanding and appreciation of prediction rule characteristics that may influence acceptance may require a qualitative study of clinicians for whom the prediction rule is intended. (297, 299)

The findings of this study suggest further research in two areas. Firstly, simplification of a prediction rule is predicated on the assumption that simplified rules, as opposed to more complex rules, are more likely to be adopted and used correctly and in entirety by clinicians. However, given that simplification may compromise the face validity of the rule which may act to prevent adoption, this assumption should be tested. Secondly, the study described one approach to the development and evaluation of simplified rules. Alternative methods of simplification and of testing performance exist, and further research is necessary to identify the most appropriate methodology.

7.4.4 Further investigation of the value of the biomarker C-reactive protein for identifying children with serious bacterial infection in the primary care setting

The findings of the review of the accuracy of C-reactive protein suggested that further research into the value of C-reactive protein as an initial diagnostic test for the workup of childhood fever would be worthwhile. Following publication of the review (in 2008), considerable

research has been undertaken, with the aim of clarifying the diagnostic accuracy and elucidating the clinical value of C-reactive protein. In the first of these studies, a review funded by the Health Technology Assessment (HTA) program in the United Kingdom, the value of C-reactive protein for ruling in and ruling out serious infection in children in ambulatory settings was examined. (222) Though the HTA review and the review presented in this thesis used slightly different methods and included some different studies, the conclusions with regard to the diagnostic accuracy of C-reactive protein were similar. More recently, research has focused on assessing the value of C-reactive protein beyond its diagnostic accuracy. In an in-process two-part clinical trial, the added value of a point of care C-reactive protein test, above and beyond readily available clinical information, will be determined in children considered to potentially have a serious bacterial infection on the basis of a clinical prediction rule. (286) Among the children without suspected serious infection (according to the prediction rule), point of care C-reactive protein testing and/or a brief intervention with safety net advice and usual care, will be evaluated for their effects on immediate antibiotic prescribing. (287)

7.5 Conclusion

The drive for a safer healthcare system has led to the development of strategies to improve the timeliness, accuracy and efficiency of diagnosis. Diagnostic prediction rules are one such strategy and, amongst developers of these tools at least, they are posited as having great potential to improve clinicians' diagnostic reasoning and clinical decision making. Ultimately, their adoption is expected to improve patient outcomes and/or the efficiency of the diagnostic process while at least maintaining patient health.

Although the clinical prediction rule literature base has expanded rapidly over recent decades, there has not been a commensurate increase in our understanding of their potential or actual ability to affect clinicians' diagnoses, decisions or to alter patient health. Further, and with perhaps some justification, clinical prediction rules have generally not been warmly welcomed or adopted by the very clinicians they seek to assist. Research in the field of diagnostic prediction rules must shift from deriving new rules to considering the use and impact of these tools in clinical practice. In such studies, patient outcomes will be the ultimate judge of whether diagnostic prediction rules offer value to the clinical diagnostic process.

Appendix A Supplementary material from the systematic review of studies comparing diagnostic clinical prediction rules with clinical judgment

PRISMA checklist (page numbers correspond to page numbers in the thesis)

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Page 37
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	Page 39
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Paragraph 2 of Introduction section, Page 40
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	Paragraph 3 of Introduction section, Page 40
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Paragraph 1 of Methods section, Page 40
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Paragraph 1 of Methods – Study Selection section, Page 41
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Paragraph 1 of Methods – Data sources and searches section, Page 41
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix A
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Paragraph 2 of Methods- Study Selection section, Page 41
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Paragraph 1 of Methods – Data extraction and risk of bias assessment section, Page 41
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Paragraph 1 of Methods – Data extraction and risk of bias assessment section, Page 41

PRISMA checklist continued

Section/topic	#	Checklist item	Reported on page #
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	Paragraph 1 of Methods –Data extraction and risk of bias assessment section, Page 41
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	Methods – Data synthesis and analysis section, Page 42
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	Methods – Data synthesis and analysis section, Page 43
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	Paragraph 1 of Methods –Data extraction and risk of bias assessment section, Page 41
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	Paragraph 2 of Methods – Data synthesis and analysis section, Page 43
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Results – Literature search section, Page 44. Figure 2.1.
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Results – Study characteristics section, Page 44 Table 2.1.
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Results – Risk of bias assessment, Page 45. Figure 2.2. Table 2.2
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Results – Study results section, Page 49-52. Table 2.3 and 2.4, Figure 2.3 and 2.4.
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Meta-analysis not conducted
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Results – Risk of bias assessment, Page 45. Figure 2.2. Table 2.2

PRISMA checklist continued.

Section/topic	#	Checklist item	Reported on page #
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	Not applicable
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	Paragraph 1 of Discussion section, Page 60
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	Paragraph 5, 6 and 7 of Discussion section, Page 62-63
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Paragraph 4, Discussion section, Page 63-64
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	Page 64

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

Electronic database search strategies

MEDLINE was searched using the Ovid interface on 24/4/13 for the period 1946 to March, Week 4, 2013	<ol style="list-style-type: none"> 1. ((clinician* or professional* or practitioner* or physician* or nurse*) adj3 (judgment* or judgement* or estimate* or diagno* or prediction* or assess* or decision* or intuition* or impression* or evaluation* or probabilit* or empirical* or subjectiv* or implicit* or unaided or unstructured or accuracy or performance)).ti,ab. 2. (clinical adj3 (judgment* or judgement* or estimate* or diagnos* or assessment* or impression* or probabilit*)).ti,ab. 3. ((empirical or subjective or implicit or unaided or unstructured) adj3 (judgment* or judgement* or estimate* or diagnos* or prediction* or assessment* or decision* or impression* or evaluation* or probabilit*)).ti,ab. 4. 1 or 2 or 3 5. *Decision Support Techniques/ 6. (scor* or rule* or model* or guide* or algorithm* or protocol* or "formal estimate" or "formal estimates").ti. 7. 5 or 6 8. 4 and 7
Embase was searched using embase.com on 27/2/12 for the period 1974 to January, 2012	<ol style="list-style-type: none"> 1. score*:ti OR rule*:ti OR model*:ti OR guide*:ti OR algorithm:ti OR protocol*:ti 2. (clinician* OR professional* OR practitioner* OR physician*) NEAR/3 (judgment OR judgement OR estimate OR diagnosis OR prediction OR assessment OR decision OR intuition OR impression OR evaluation OR probability OR empirical OR subjective OR implicit OR unaided OR unstructured) 3.(empirical OR subjective OR implicit OR unaided) NEAR/3 (judgment OR judgement OR estimate OR diagnosis OR prediction OR assessment OR decision OR impression OR evaluation OR probability) 4. 2 or 3 5. 1 and 3
Cumulative Index to Nursing and Allied Health Literature (CINAHL) using the EBSCOhost interface on 27/2/12 for the period 1982 to January, 2012	<ol style="list-style-type: none"> 1. TI (score* OR rule* OR model* OR guide* OR algorithm OR protocol OR "formal estimate") 2. TI (((clinician* OR clinical OR professional* OR practitioner* OR physician* OR nurse*) N3 (judgment OR estimate OR diagnosis OR prediction OR assessment OR decision OR intuition OR impression OR evaluation OR probability OR empirical OR subjective OR implicit OR unaided))) OR AB (((clinician* OR clinical OR professional* OR practitioner* OR physician* OR nurse*) N3 (judgement OR judgment OR estimate OR diagnosis OR prediction OR assessment OR decision OR intuition OR impression OR evaluation OR probability OR empirical OR subjective OR implicit OR unaided))) 3. TI (((clinical) N3 (judgment OR estimate OR diagnosis OR assessment OR impression OR probability))) OR AB (((clinical) N3 (judgment OR estimate OR diagnosis OR assessment OR impression OR probability))) 4. TI (((empirical OR subjective OR implicit OR unaided) N3 (judgment OR judgement OR estimate OR diagnosis OR prediction OR assessment OR decision OR impression OR evaluation OR probability))) OR (((empirical OR subjective OR implicit OR unaided) N3 (judgment OR estimate OR diagnosis OR prediction OR assessment OR decision OR impression OR evaluation OR probability))) 5. 2 or 3 or 4 6. 1 and 5
PubMed was searched on 28/4/13 for systematic reviews of clinical prediction rules	<ol style="list-style-type: none"> 1. Medline[tiab] OR (systematic[tiab] AND review[tiab]) OR meta-analysis[ptyp] 2. Score[ti] OR scores[ti] OR rule[ti] OR rules[ti] 3. 1 AND 2

Appendix B Supplementary material from the systematic review of the effects of diagnostic clinical prediction rules

PRISMA checklist (page numbers correspond to page numbers in the thesis)

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Page 65
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	Page 67
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Paragraph 3 of Introduction section, Page 68
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	Paragraph 4 of Introduction section, Page 68
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Paragraph 1 of Methods section, Page 69
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Paragraph 1 of Methods – Study Selection section, Page 69
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Paragraph 1 of Methods – Data sources and searches section, Page 69
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix B
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Paragraph 3 of Methods- Study Selection section, Page 69-70
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Paragraph 1 and of Methods – Data extraction, risk of bias and data synthesis section, Page 70
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Paragraph 2 and 3 of Methods – Data extraction, risk of bias and data synthesis section, Page 70

PRISMA checklist continued

Section/topic	#	Checklist item	Reported on page #
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	Paragraph 1 of Methods –Risk of bias assessment section, Page 71
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	Methods – Data synthesis section, Page 71
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	Methods – Data synthesis section, Page 71-72
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	Paragraph 1 of Risk of bias assessment section, Page 71
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	Paragraph 2 of Data synthesis section, Page 71-72
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Results – Study selection section, Page 72. Figure 3.1 Page 75.
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Results – Trial characteristics section, Page 73-74. Table 3.1.
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Results – Risk of bias. Page 80. Table 3.2. Page
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Results – Effects of diagnostic strategies incorporating diagnostic clinical prediction rules section, Page 82-93 Table 3.3 to 3.8.
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Results – Effects of diagnostic strategies incorporating diagnostic clinical prediction rules section, Page 84 and 88. Figures 3.2 and 3.3.
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Results – Risk of bias. Page 80

PRISMA checklist continued

Section/topic	#	Checklist item	Reported on page #
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	Results- Studies of acute appendicitis – Clinicians' decisions section. Page 85. Results – Assessment of reporting of interventions section. Page 94. Table 3.9 Page 95-96.
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	Paragraph 1 of Discussion section, Page 97
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	Paragraph 2, 4, 7 and 10 of Discussion section, Page 97-98, 101.
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Paragraph 1 of Conclusions section, Page 102
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	Page 102

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097.
doi:10.1371/journal.pmed1000097

Electronic database search strategies

<p>Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) 1946 to 30th May 2015. Searched on 16/06/2015.</p>	<ol style="list-style-type: none"> 1. Randomized controlled trial.pt. 2. Controlled clinical trial.pt. 3. Randomized.ab. 4. Placebo.ab. 5. Clinical trials as topic.sh. 6. Randomly.ab. 7. Trial.ti. 8. 1 or 2 or 3 or 4 or 5 or 6 or 7 9. exp animals/ not humans.sh. 10. 8 not 9 11. *Decision Support Systems, Clinical/ 12. Rule*.ti. 13. (rule* adj3 (decision OR clinical OR diagnos* OR predict*).ti,ab. 14. Score*.ti. 15. (score* adj3 (decision OR clinical OR diagnos* OR predict* OR risk).ti,ab. 16. ((aid or model) adj3 (clinical or decision or diagnos* or predict*).ti,ab. 17. ((guide or algorithm or protocol) adj2 (diagnos* or decision or clinical or predict*).ti,ab. 18. 11 or 12 or 13 or 14 or 15 or 16 or 17 19. 18 and 10
<p>Cochrane Central Register of Controlled Trials (CENTRAL, Issue 5 of 12 May 2015) in the Cochrane Library. Searched 24/6/2015</p>	<p>(score* OR rule* OR ((protocol OR aid OR algorithm OR tool OR instrument) near/2 (diagnos* OR decision OR clinical OR predict*))):ti</p>

Table of studies excluded from the review with reasons for exclusion

Reference	Comment
Not a study of randomised allocation to care with and without a prediction rule	
Ackerman SL, Gonzales R, Stahl MS, Metlay JP. One size does not fit all: evaluating an intervention to reduce antibiotic prescribing for acute bronchitis. BMC Health Serv Res. 2013;13:462. PubMed PMID: 24188573. Pubmed Central PMCID: 4228248.	Not a randomised comparison
Adams ID, Chan M, Clifford PC, Cooke WM, Dallos V, Dombal FT, et al. Computer aided diagnosis of acute abdominal pain: a multicentre study. British Medical Journal [Internet]. 1986; 293(6550):[800-4 pp.].	Before after study
Ammirati F, Colivicchi F, Santini M. Diagnosing syncope in clinical practice. Implementation of a simplified diagnostic algorithm in a multicentre prospective trial - the OESIL 2 study (Osservatorio Epidemiologico della Sincope nel Lazio). European Heart Journal. 2000;21(11):935-40. PubMed PMID: 10806018.	Not a controlled study
Beltrán MA, Villar MR, Cruces KS. [Application of a diagnostic score for appendicitis by health-related non-physician professionals]. Revista médica de Chile [Internet]. 2006; 134(1):[39-47 pp.]	Not a randomised comparison
Bajaj RR, Goodman SG, Yan RT, Bagnall AJ, Gyenes G, Welsh RC, et al. Treatment and outcomes of patients with suspected acute coronary syndromes in relation to initial diagnostic impressions (insights from the Canadian Global Registry of Acute Coronary Events [GRACE] and Canadian Registry of Acute Coronary Events [CANRACE]). American Journal of Cardiology. 2013;111(2):202-7. PubMed PMID: 23122889.	Not a randomised comparison
Bessen T, Clark R, Shakib S, Hughes G. A multifaceted strategy for implementation of the Ottawa ankle rules in two emergency departments. BMJ. 2009;339:b3056. PubMed PMID: 19675080. Pubmed Central PMCID: 2726279.	Before after study
Bressan S, editor Implementation of PECARN decision rule for children with minor head injury in the pediatric emergency department. Mediterranean Emergency Medicine Congress (VI); 2011.	Before after study
Boutis K, Grootendorst P, Willan A, Plint AC, Babyn P, Brison RJ, et al. Effect of the Low Risk Ankle Rule on the frequency of radiography in children with ankle injuries. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2013 Oct 15;185(15):E731-8. PubMed PMID: 23939215. Pubmed Central PMCID: 3796622.	Not a randomised comparison. An interrupted time series study with matched control
Brand DA, Frazier WH, Kohlhepp WC, Shea KM, Hoefer AM, Ecker MD, et al. A protocol for selecting patients with injured extremities who need x-rays. N Engl J Med. 1982 Feb 11;306(6):333-9. PubMed PMID: 7054709.	Not a controlled study
Broekhuizen BD, Sachs A, Janssen K, Geersing GJ, Moons K, Hoes A, et al. Does a decision aid help physicians to detect chronic obstructive pulmonary disease? The British journal of general practice : the journal of the Royal College of General Practitioners. 2011 Oct;61(591):e674-9. PubMed PMID: 22152850. Pubmed Central PMCID: 3177137.	Not a randomised comparison. Study of the incremental value of the prediction rule
Cameron C, Naylor CD. No impact from active dissemination of the Ottawa Ankle Rules: further evidence of the need for local implementation of practice guidelines. Cmaj [Internet]. 1999; 160(8):[1165-8 pp.].	Not a randomised comparison. Before after study of the effect of an active dissemination strategy. Investigators compared use of ankle radiography in hospitals receiving an educational intervention with some or no use of the OARs with hospitals who declined the dissemination strategy (already using the OARs).
Casey JR, Block S, Puthoor P, Hedrick J, Almudevar A, Pichichero ME. A simple scoring system to improve clinical assessment of acute otitis media. Clinical pediatrics [Internet]. 2011; 50(7):[623-9 pp.].	Not a controlled study
Christian F, Christian GP. A simple scoring system to reduce the negative appendectomy rate. Annals of the Royal College of Surgeons of England [Internet]. 1992; 74(4):[281-5 pp.].	Not a randomised comparison
Courtney DM, Kline JA. Prospective use of a clinical decision rule to identify pulmonary embolism as likely cause of outpatient cardiac arrest. Resuscitation [Internet]. 2005; 65(1):[57-64 pp.].	Not a controlled study
den Exter PL, Gomez V, Jimenez D, Trujillo-Santos J, Muriel A, Huisman MV, et al. A clinical prognostic model for the identification of low-risk patients with acute symptomatic pulmonary embolism and active cancer. Chest. 2013;143(1):138-45. PubMed PMID: 22814859.	Not a controlled study
Dobbs F. A scoring system for predicting group A streptococcal throat infection. British Journal of General Practice. 1996;46(409):461-4. PubMed PMID: 8949324. Pubmed Central PMCID: PMC1239715.	Not a controlled study
Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. British medical journal [Internet]. 1972; 2(5804):[9-13 pp.].	Not a controlled study
Drescher FS, Chandrika S, Weir ID, Weintraub JT, Berman L, Lee R, et al. Effectiveness and acceptability of a computerized decision support system using modified Wells criteria for evaluation of suspected pulmonary embolism. Ann Emerg Med. 2011 Jun;57(6):613-21. PubMed PMID: 21050624.	Not a randomised comparison
Enochsson L, Gudbjartsson T, Hellberg A, Rudberg C, Wenner J, Ringqvist I, et al. The Fenyö-Lindberg scoring system for appendicitis increases positive predictive value in fertile women--a prospective study in 455 patients randomized to either laparoscopic or open appendectomy. Surgical endoscopy [Internet]. 2004; 18(10):[1509-13 pp.].	Not a controlled study
Eccles M, Steen N, Grimshaw J, Thomas L, McNamee P, Soutter J, et al. Effect of audit and feedback,	No control group without Guideline.Guideline does

and reminder messages on primary-care radiology referrals: a randomised trial. <i>Lancet</i> . 2001 May 5;357(9266):1406-9. PubMed PMID: 11356439.	not appear to include a CPR.
Gonzales R, Aagaard EM, Camargo CA, Jr., Ma OJ, Plautz M, Maselli JH, et al. C-reactive protein testing does not decrease antibiotic use for acute cough illness when compared to a clinical algorithm. <i>Journal of Emergency Medicine</i> . 2011;41(1):1-7. PubMed PMID: 19095403.	No control group without CPR
Green L, Mehr DR. What alters physicians' decisions to admit to the coronary care unit? <i>J Fam Pract</i> . 1997 Sep;45(3):219-26. PubMed PMID: 9300001.	Interrupted time series
Goldman L, Cook EF, Brand DA, Lee TH, Rouan GW, Weisberg MC, et al. A computer protocol to predict myocardial infarction in emergency department patients with chest pain. <i>N Engl J Med</i> . 1988 Mar 31;318(13):797-803. PubMed PMID: 3280998.	Not a controlled study
Holroyd BR, Wilson D, Rowe BH, Mayes DC, Noseworthy T. Uptake of validated clinical practice guidelines: experience with implementing the Ottawa Ankle Rules. <i>The American journal of emergency medicine [Internet]</i> . 2004; 22(3):[149-55 pp.].	Not a randomised comparison
Harizman N, Oliveira C, Chiang A, Tello C, Marmor M, Ritch R, et al. The ISNT rule and differentiation of normal from glaucomatous eyes. <i>Archives of ophthalmology [Internet]</i> . 2006; 124(11):[1579-83 pp.].	Not a controlled study
Hsu P, Lam LT, Browne G. The pulmonary index score as a clinical assessment tool for acute childhood asthma. <i>Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology [Internet]</i> . 2010; 105(6):[425-9 pp.].	Not a randomised comparison
Iapichino G, Mistraretti G, Corbella D, Bassi G, Borotto E, Miranda DR, et al. Scoring system for the selection of high-risk patients in the intensive care unit. <i>Critical care medicine [Internet]</i> . 2006; 34(4):[1039-43 pp.].	Not a randomised comparison
Jacobs AA. Clinical prediction of pneumonia. <i>Annals of Internal Medicine</i> . 1991;114(4):428.	Letter to editor
Kerr D, Bradshaw L, Kelly AM. Implementation of the Canadian C-spine rule reduces cervical spine x-ray rate for alert patients with potential neck injury. <i>J Emerg Med</i> . 2005 Feb;28(2):127-31. PubMed PMID: 15707805.	Not a randomised comparison
Kilroy DA, Ireland S, Reid P, Goodacre S, Morris F. Emergency department investigation of deep vein thrombosis. <i>Emergency medicine journal : EMJ</i> . 2003 Jan;20(1):29-32. PubMed PMID: 12533363. Pubmed Central PMCID: 1726005.	Not a controlled study
Kec RM, Richman PB, Szucs PA, Mandell M, Eskin B. Can emergency department triage nurses appropriately utilize the Ottawa Knee Rules to order radiographs?-An implementation trial. <i>Academic Emergency Medicine</i> . 2003;10(2):146-50. PubMed PMID: 12574012.	Not a controlled study
Khan I, Rehman AU. Application of alvarado scoring system in diagnosis of acute appendicitis. <i>Journal of Ayub Medical College Abbottabad [Internet]</i> . 2005; 17(3):[41-4 pp.].	Not a controlled study
Kotowycz MA, Cosman TL, Tartaglia C, Afzal R, Syal RP, Natarajan MK. Safety and feasibility of early hospital discharge in ST-segment elevation myocardial infarction--a prospective and randomized trial in low-risk primary percutaneous coronary intervention patients (the Safe-Depart Trial). <i>American Heart Journal</i> . 2010;159(1):117.e1-6. PubMed PMID: 20102876.	Not a randomised comparison (diagnostic CPR used for assessing inclusion in the study)
Lopez PP, Cohn SM, Popkin CA, Jackowski J, Michalek JE, Appendicitis Diagnostic G. The use of a computed tomography scan to rule out appendicitis in women of childbearing age is as accurate as clinical examination: a prospective randomized trial. <i>American Surgeon</i> . 2007;73(12):1232-6. PubMed PMID: 18186378.	Not a randomised comparison (diagnostic CPR used for assessing inclusion in the study)
Leddy JJ, Kesari A, Smolinski RJ. Implementation of the Ottawa ankle rule in a university sports medicine centre. <i>Medicine and science in sports and exercise</i> . 2002 Jan;34(1):57-62. PubMed PMID: 11782648.	Not a controlled study
Lee TH, Pearson SD, Johnson PA, Garcia TB, Weisberg MC, Guadagnoli E, et al. Failure of information as an intervention to modify clinical management. A time-series trial in patients with acute chest pain. <i>Ann Intern Med</i> . 1995 Mar 15;122(6):434-7. PubMed PMID: 7856992.	Not a randomised comparison
Matloob SA, Roach J, Marcus HJ, O'Neill K, Nair R. Evaluation of the impact of the Canadian subarachnoid haemorrhage clinical decision rules on British practice. <i>British journal of neurosurgery</i> . 2013 Oct;27(5):603-6. PubMed PMID: 23730979.	Not a controlled study
McAdam WA, Brock BM, Armitage T, Davenport P, Chan M, de Dombal FT. Twelve years' experience of computer-aided diagnosis in a district general hospital. <i>Ann R Coll Surg Engl</i> . 1990 Mar;72(2):140-6. PubMed PMID: 2185682. Pubmed Central PMCID: 2499113.	Interrupted time series study
McIsaac WJ, Kellner JD, Aufricht P, Vanjaka A, Low DE. Empirical validation of guidelines for the management of pharyngitis in children and adults. <i>JAMA</i> . 2004 Apr 7;291(13):1587-95. PubMed PMID: 15069046.	Not a randomised comparison
Mortola GP, Arnulfo G, Reboa G, Pitto G, Masini R, DiSomma C, et al. Clinical application of a computerized diagnostic aid in the initial evaluation of 250 outpatients with gastrointestinal complaints. <i>Journal of Clinical Computing</i> . 1987;16(3-4):93-103. PubMed PMID: 10302545.	Not a randomised comparison
Mainous AG, 3rd, Lambourne CA, Nietert PJ. Impact of a clinical decision support system on antibiotic prescribing for acute respiratory infections in primary care: quasi-experimental trial. <i>Journal of the American Medical Informatics Association</i> . 2013;20(2):317-24. PubMed PMID: 22759620. Pubmed Central PMCID: PMC3638170	Not a randomised comparison
Man E, Simonka Z, Varga A, Raros F, Lazar G. Impact of the Alvarado score on the diagnosis of acute appendicitis: comparing clinical judgment, Alvarado score, and a new modified score in suspected appendicitis: a prospective, randomized clinical trial. <i>Surg Endosc</i> . 2014 Aug;28(8):2398-405. PubMed PMID: 24705731.	Not a randomised comparison
Naschitz JE, Rosner I, Rozenbaum M, Naschitz S, Musafia-Priselac R, Shaviv N, et al. The head-up	Not a controlled study

tilt test with haemodynamic instability score in diagnosing chronic fatigue syndrome. QJM : monthly journal of the Association of Physicians [Internet]. 2003; 96(2):[133-42 pp.]. Available from: http://onlinelibrary.wiley.com/o/cochrane/clcentral/articles/734/CN-00422734/frame.html .	
Ohmann C, Franke C, Yang Q. Clinical benefit of a diagnostic score for appendicitis: results of a prospective interventional study. German Study Group of Acute Abdominal Pain. Arch Surg. 1999 Sep;134(9):993-6. PubMed PMID: 10487595.	Before after study
Owen TD, Williams H, Stiff G, Jenkinson LR, Rees BI. Evaluation of the Alvarado score in acute appendicitis. J R Soc Med. 1992 Feb;85(2):87-8. PubMed PMID: 1489366. Pubmed Central PMCID: 1294889.	Not a randomised comparison
Pitt E, Pedley DK, Nelson A, Cumming M, Johnston M. Removal of C-spine protection by A&E triage nurses: a prospective trial of a clinical decision making instrument. Emergency Medicine Journal. 2006;23(3):214-5. PubMed PMID: 16498160. Pubmed Central PMCID: PMC2464447.	Not a controlled study
Poses RM, Cebul RD, Wigton RS, Centor RM, Collins M, Fleischli G. Controlled trial using computerized feedback to improve physicians' diagnostic judgments. Academic Medicine. 1992;67(5):345-7. PubMed PMID: 1575873.	Non randomised before after study with concurrent controls
Pozen MW, D'Agostino RB, Mitchell JB, Rosenfeld DM, Guglielmino JT, Schwartz ML, et al. The usefulness of a predictive instrument to reduce inappropriate admissions to the coronary care unit. Annals of internal medicine [Internet]. 1980; 92(2 Pt 1):[238-42 pp.].	Interrupted time series
Pozen MW, D'Agostino RB, Selker HP, Sytkowski PA, Hood WB, Jr. A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease. A prospective multicenter clinical trial. New England Journal of Medicine. 1984;310(20):1273-8. PubMed PMID: 6371525.	Not a randomised comparison (allocated by alternate months or 6 month period)/reference standard for diagnosis not current
Roy PM, Durieux P, Gillaizeau F, Legall C, Armand-Perroux A, Martino L, et al. A computerized handheld decision-support system to improve pulmonary embolism diagnosis: a randomized trial. Annals of Internal Medicine. 2009;151(10):677-86. PubMed PMID: 19920268.	Before after study
Roberts RR, Zalenski RJ, Mensah EK, Rydman RJ, Ciavarella G, Gussow L, et al. Costs of an emergency department-based accelerated diagnostic protocol vs hospitalization in patients with chest pain: a randomized controlled trial. JAMA. 1997;278(20):1670-6. PubMed PMID: 9388086.	Not a controlled study
Reilly BM, Evans AT, Schaidler JJ, Das K, Calvin JE, Moran LA, et al. Impact of a clinical decision rule on hospital triage of patients with suspected acute cardiac ischemia in the emergency department. JAMA. 2002 Jul 17;288(3):342-50. PubMed PMID: 12117399.	Interrupted time series
Righini M, Le Gal G, Aujesky D, Roy PM, Sanchez O, Verschuren F, et al. Diagnosis of pulmonary embolism by multidetector CT alone or combined with venous ultrasonography of the leg: a randomised non-inferiority trial. Lancet. 2008 Apr 19;371(9621):1343-52. PubMed PMID: 18424324.	Not a randomised comparison of CPR vs no CPR. CPR used in all participants who were then randomised to different sequences of testing
Reilly BM, Evans AT, Schaidler JJ, Wang Y. Triage of patients with chest pain in the emergency department: a comparative study of physicians' decisions. Am J Med. 2002 Feb 1;112(2):95-103. PubMed PMID: 11835946.	Not a randomised comparison
Sarasin FP, Raymond JM, Griffith JL, Beshansky JR, Schifferli JA, Unger PF, et al. Impact of the acute cardiac ischemia time-insensitive predictive instrument (ACI-TIPI) on the speed of triage decision making for emergency department patients presenting with chest pain: a controlled clinical trial. Journal of general internal medicine [Internet]. 1994; 9(4):[187-94 pp.].	No control group without CPR
Scheye T, Vanneuville G. [Trial of a diagnostic score in painful abdominal syndromes suggestive of appendicitis in children over 3]. Journal de Chirurgie. 1988;125(3):166-9. PubMed PMID: 3372603.	Not a randomised comparison (diagnostic CPR used for assessing inclusion in the study)
Selker HP, Beshansky JR, Griffith JL. Use of the electrocardiograph-based thrombolytic predictive instrument to assist thrombolytic and reperfusion therapy for acute myocardial infarction. A multicenter, randomized, controlled, clinical effectiveness trial. Annals of internal medicine [Internet]. 2002; 137(2):[87-95 pp.].	Interrupted time series
Selker HP, Beshansky JR, Griffith JL, Aufderheide TP, Ballin DS, Bernard SA, et al. Use of the acute cardiac ischemia time-insensitive predictive instrument (ACI-TIPI) to assist with triage of patients with chest pain or other symptoms suggestive of acute cardiac ischemia. A multicenter, controlled clinical trial. Annals of Internal Medicine. 1998;129(11):845-55. PubMed PMID: 9867725.	Not a controlled study
Selker HP, Beshansky JR, Ruthazer R, Sheehan PR, Sayah AJ, Atkins JM, et al. Emergency medical service predictive instrument-aided diagnosis and treatment of acute coronary syndromes and ST-segment elevation myocardial infarction in the IMMEDIATE trial. Prehospital Emergency Care. 2011;15(2):139-48. PubMed PMID: 21366431.	Not a randomised comparison
Stiell I, Wells G, Laupacis A, Brison R, Verbeek R, Vandemheen K, et al. Multicentre trial to introduce the Ottawa ankle rules for use of radiography in acute ankle injuries. Multicentre Ankle Rule Study Group. BMJ. 1995;311(7005):594-7. PubMed PMID: 7663253. Pubmed Central PMCID: PMC2550661.	Interrupted time series
Sutton GC. How accurate is computer-aided diagnosis? Lancet [Internet]. 1989; 2(8668):[905-8 pp.].	Not a randomised comparison
Stiell IG, McKnight RD, Greenberg GH, McDowell I, Nair RC, Wells GA, et al. Implementation of the Ottawa ankle rules. JAMA. 1994 Mar 16;271(11):827-32. PubMed PMID: 8114236.	Non randomised before after study with concurrent controls
Stiell IG, Wells GA, Hoag RH, Sivilotti ML, Cacciotti TF, Verbeek PR, et al. Implementation of the Ottawa Knee Rule for the use of radiography in acute knee injuries. JAMA. 1997;278(23):2075-9. PubMed PMID: 9403421.	Non randomised before after study with concurrent controls
Singh S, Nosyk B, Sun H, Christenson JM, Innes G, Anis AH. Value of information of a clinical prediction rule: informing the efficient use of healthcare and health research resources. Int J Technol Assess Health Care. 2008 Winter;24(1):112-9. PubMed PMID: 18218176.	Decision analytic modelling study

Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C, et al. A simple clinical model for the diagnosis of deep-vein thrombosis combined with impedance plethysmography: potential for an improvement in the diagnostic process. <i>J Intern Med</i> . 1998 Jan;243(1):15-23. PubMed PMID: 9487327.	Not a controlled study
Woolley SL, Bernstein JM, Davidson JA, Smith DR. Sore throat in adults--does the introduction of a clinical scoring system improve the management of these patients in a secondary care setting? <i>The Journal of laryngology and otology</i> . 2005 Jul;119(7):550-5. PubMed PMID: 16175981.	Before after study
Weingarten S, Ermann B, Bolus R, Riedinger MS, Rubin H, Green A, et al. Early "step-down" transfer of low-risk patients with chest pain. A controlled interventional trial. <i>Annals of Internal Medicine</i> . 1990;113(4):283-9. PubMed PMID: 2115754.	
Wells PS, Anderson DR, Rodger M, Forgie M, Kearon C, Dreyer J, et al. Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis. <i>New England Journal of Medicine</i> . 2003;349(13):1227-35. PubMed PMID: 14507948.	Not a randomised comparison of CPR versus no CPR (CPR applied to all then randomised to different management strategy based on output of CPR)
Wilson PD, Horrocks JC, Lyndon PJ, Yeung CK, Page RE, Dombal FT. Simplified computer-aided diagnosis of acute abdominal pain. <i>British medical journal [Internet]</i> . 1975; 2(5962):[73-5 pp.].	Not a controlled study
Winn RD, Laura S, Douglas C, Davidson P, Gani JS. Protocol-based approach to suspected appendicitis, incorporating the Alvarado score and outpatient antibiotics. <i>ANZ J Surg</i> . 2004 May;74(5):324-9. PubMed PMID: 15144250.	Not a randomised comparison
Wilson EC, Emery JD, Kinmonth AL, Prevost AT, Morris HC, Humphrys E, et al. The cost-effectiveness of a novel SIAscopic diagnostic aid for the management of pigmented skin lesions in primary care: a decision-analytic model. <i>Value in Health</i> . 2013;16(2):356-66. PubMed PMID: 23538188.	Decision analytic modelling economic study incorporating information from RCT included in the review
Westfall JM, Van Vorst RF, McGloin J, Selker HP. Triage and diagnosis of chest pain in rural hospitals: implementation of the ACI-TIPI in the High Plains Research Network. <i>Ann Fam Med</i> . 2006 Mar-Apr;4(2):153-8. PubMed PMID: 16569719. Pubmed Central PMCID: 1467005.	Not a randomised comparison
Intervention does not include a diagnostic prediction rule or the prediction rule is intended to facilitate clinician and patient shared decision making	
Achaval S, Fraenkel L, Volk RJ, Cox V, Suarez-Almazor ME. Impact of educational and patient decision aids on decisional conflict associated with total knee arthroplasty. <i>Arthritis care & research [Internet]</i> . 2012; 64(2):[229-37 pp.].	Intervention is a patient decision aid not a diagnostic CPR for clinician use only
Bell LM, Grundmeier R, Localio R, Zorc J, Fiks AG, Zhang X, et al. Electronic health record-based decision support to improve asthma care: a cluster-randomized trial. <i>Pediatrics</i> . 2010;125(4):e770-7. PubMed PMID: 20231191.	Intervention does not include a diagnostic CPR
Bates DW, Kuperman GJ, Rittenberg E, Teich JM, Fiskio J, Ma'luf N, et al. A randomized trial of a computer-based intervention to reduce utilization of redundant laboratory tests. <i>Am J Med</i> . 1999 Feb;106(2):144-50. PubMed PMID: 10230742.	Intervention does not include a diagnostic CPR
Bourgeois FC, Linder J, Johnson SA, Co JP, Fiskio J, Ferris TG. Impact of a computerized template on antibiotic prescribing for acute respiratory infections in children and adolescents. <i>Clin Pediatr (Phila)</i> . 2010 Oct;49(10):976-83. PubMed PMID: 20724348.	Intervention does not include a diagnostic CPR
Christakis DA, Zimmerman FJ, Wright JA, Garrison MM, Rivara FP, Davis RL. A randomized controlled trial of point-of-care evidence to improve the antibiotic prescribing practices for otitis media in children. <i>Pediatrics</i> . 2001;107(2):E15. PubMed PMID: 11158489.	Intervention does not include a diagnostic CPR
Carroll AE, Biondich P, Anand V, Dugan TM, Downs SM. A randomized controlled trial of screening for maternal depression with a clinical decision support system. <i>Journal of the American Medical Informatics Association</i> . 2013;20(2):311-6. PubMed PMID: 22744960.	Intervention does not include a diagnostic CPR
Chang AB, Robertson CF, van Asperen PP, Glasgow NJ, Masters IB, Teoh L, et al. A cough algorithm for chronic cough in children: a multicenter, randomized controlled study. <i>Pediatrics</i> . 2013;131(5):e1576-83. PubMed PMID: 23610200.	Intervention does not include a diagnostic CPR
Cheyne H, Hundley V, Dowding D, Bland JM, McNamee P, Greer I, et al. Effects of algorithm for diagnosis of active labour: cluster randomised trial. <i>BMJ (Clinical research ed) [Internet]</i> . 2008; 337:[a2396 p.].	Intervention does not include a diagnostic CPR as defined by review (predictors identified by expert opinion and literature review not multivariable statistical analysis)
Clayton TC, Lubsen J, Pocock SJ, Vokó Z, Kirwan BA, Fox KA, et al. Risk score for predicting death, myoc and stroke in patients with stable angina, based on a large randomised trial cohort of patients. <i>BMJ (Cl [Internet]</i> . 2005; 331(7521):[869 p.].	Intervention does not include a diagnostic CPR
Carroll AE, Bauer NS, Dugan TM, Anand V, Saha C, Downs SM. Use of a computerized decision aid for ADHD diagnosis: a randomized controlled trial. <i>Pediatrics</i> . 2013 Sep;132(3):e623-9. PubMed PMID: 23958768. Pubmed Central PMCID: 3876764.	Intervention does not include a diagnostic CPR
Chang AB, Robertson CF, van Asperen PP, Glasgow NJ, Masters IB, Teoh L, et al. A cough algorithm for chronic cough in children: a multicenter, randomized controlled study. <i>Pediatrics</i> . 2013;131(5):e1576-83. PubMed PMID: 23610200.	Intervention does not include a diagnostic CPR
Coutinho Storti F, Moffa PJ, Uchida AH, Hueb WA, Machado Cesar LA, Ferreira BM, et al. New prognostic score for stable coronary disease evaluation. <i>Arquivos Brasileiros de Cardiologia</i> . 2011;96(5):411-8. PubMed PMID: 21503388.	Intervention does not include a diagnostic CPR
Del Mar CB, Green AC. Aid to diagnosis of melanoma in primary medical care. <i>BMJ</i> . 1995;310(6978):492-5. PubMed PMID: 7888887. Pubmed Central PMCID: PMC2548872.	Intervention does not include a diagnostic CPR as defined by review (predictors identified by expert opinion not multivariable statistical analysis)

Dexheimer JW, Abramo TJ, Arnold DH, Johnson K, Shyr Y, Ye F, et al. Implementation and evaluation of an integrated computerized asthma management system in a pediatric emergency department: a randomized clinical trial. <i>Int J Med Inform.</i> 2014 Nov;83(11):805-13. PubMed PMID: 25174321.	Intervention does not incorporate a diagnostic CPR
Emmett CL, Montgomery AA, Peters TJ, Fahey T. Three-year follow-up of a factorial randomised controlled trial of two decision aids for newly diagnosed hypertensive patients. <i>British Journal of General Practice [Internet]</i> . 2005; 55(516):[551-3 pp.].	Intervention does not include a diagnostic CPR
English DR, Burton RC, del Mar CB, Donovan RJ, Ireland PD, Emery G. Evaluation of aid to diagnosis of pigmented skin lesions in general practice: controlled trial randomised by practice. <i>BMJ.</i> 2003;327(7411):375. PubMed PMID: 12919990. Pubmed Central PMCID: PMC175808.	Intervention does not include a diagnostic CPR as defined by review (uses CPR of Del Mar 1995 above)
Forrest CB, Fiks AG, Bailey LC, Localio R, Grundmeier RW, Richards T, et al. Improving adherence to otitis media guidelines with clinical decision support and physician feedback. <i>Pediatrics.</i> 2013;131(4):e1071-81. PubMed PMID: 23478860.	Intervention does not include a diagnostic CPR
Foy R, Penney GC, Grimshaw JM, Ramsay CR, Walker AE, MacLennan G, et al. A randomised controlled trial of a tailored multifaceted strategy to promote implementation of a clinical guideline on induced abortion care. <i>BJOG.</i> 2004 Jul;111(7):726-33. PubMed PMID: 15198764.	Intervention does not incorporate a diagnostic CPR
Gonzales R, Anderer T, McCulloch CE, Maselli JH, Bloom FJ, Jr., Graf TR, et al. A cluster randomized trial of decision support strategies for reducing antibiotic use in acute bronchitis. <i>JAMA Intern Med.</i> 2013 Feb 25;173(4):267-73. PubMed PMID: 23319069. Pubmed Central PMCID: 3582762.	Intervention does not include a diagnostic CPR
Hagiwara M, Henricson M, Jonsson A, Suserud BO. Decision-support tool in prehospital care: a systematic review of randomized trials. <i>Prehospital & Disaster Medicine.</i> 2011;26(5):319-29. PubMed PMID: 22030101.	Intervention does not include a diagnostic CPR
Hess EP, Knoedler MA, Shah ND, Kline JA, Breslin M, Branda ME, et al. The chest pain choice decision aid: a randomized trial. <i>Circulation Cardiovascular Quality & Outcomes.</i> 2012;5(3):251-9. PubMed PMID: 22496116.	Intervention is a patient decision aid not a diagnostic CPR for clinician use only
Hamilton E, Platt R, Gauthier R, McNamara H, Miner L, Rothenberg S, et al. The effect of computer-assisted evaluation of labor on cesarean rates. <i>Journal for healthcare quality : official publication of the National Association for Healthcare Quality.</i> 2004 Jan-Feb;26(1):37-44. PubMed PMID: 14763319.	Intervention does not include a diagnostic CPR as defined by review
Hoffmann U, Truong QA, Fleg JL, Goehler A, Gazelle S, Wiviott S, et al. Design of the Rule Out Myocardial Ischemia/Infarction Using Computer Assisted Tomography: a multicenter randomized comparative effectiveness trial of cardiac computed tomography versus alternative triage strategies in patients with acute chest pain in the emergency department. <i>American heart journal [Internet]</i> . 2012; 163(3):[330-8, 8.e1 pp.].	Intervention does not include a diagnostic CPR
Kucher N, Koo S, Quiroz R, Cooper JM, Paterno MD, Soukonnikov B, et al. Electronic alerts to prevent venous thromboembolism among hospitalized patients. <i>New England Journal of Medicine.</i> 2005;352(10):969-77. PubMed PMID: 15758007.	Intervention does not include a diagnostic CPR developed in other data
Kline JA, Zeitouni RA, Hernandez-Nino J, Jones AE. Randomized trial of computerized quantitative pretest probability in low-risk chest pain patients: effect on safety and resource use. <i>Ann Emerg Med.</i> 2009 Jun;53(6):727-35 e1. PubMed PMID: 19135281.	Study of a patient decision aid not a diagnostic CPR for clinician use only
Lindley-Jones M, Finlayson BJ. Triage nurse requested x rays--are they worthwhile? <i>J Accid Emerg Med.</i> 2000 Mar;17(2):103-7. PubMed PMID: 10718230. Pubmed Central PMCID: 1725357.	Intervention does not include a diagnostic CPR
Llor C, Madurell J, Balague-Corbella M, Gomez M, Cots JM. Impact on antibiotic prescription of rapid antigen detection testing in acute pharyngitis in adults: a randomised clinical trial. <i>Br J Gen Pract.</i> 2011 May;61(586):e244-51. PubMed PMID: 21619748. Pubmed Central PMCID: 3080229.	Intervention does not include a diagnostic CPR
Lewis G, Sharp D, Bartholomew J, Pelosi AJ. Computerized assessment of common mental disorders in primary care: effect on clinical outcome. <i>Fam Pract.</i> 1996 Apr;13(2):120-6. PubMed PMID: 8732321.	Intervention does not include a diagnostic CPR
Lee NJ, Chen ES, Currie LM, Donovan M, Hall EK, Jia H, et al. The effect of a mobile clinical decision support system on the diagnosis of obesity and overweight in acute and primary care encounters. <i>Advances in Nursing Science.</i> 2009;32(3):211-21. PubMed PMID: 19707090.	Intervention does not include a diagnostic CPR
Lironi A, Zawadzinski S, La Scala G, Thevenod C, Le Coultre C. [Value of the Pediatric Trauma Score in routine hospital practice--apropos of a prospective one-year trial]. <i>Swiss Surgery.</i> 1999;5(6):271-5. PubMed PMID: 10608189.	Intervention does not include a diagnostic CPR
Marrie TJ, Lau CY, Wheeler SL, Wong CJ, Vandervoort MK, Feagan BG. A controlled trial of a critical pathway for treatment of community-acquired pneumonia. CAPITAL Study Investigators. Community-Acquired Pneumonia Intervention Trial Assessing Levofloxacin. <i>JAMA.</i> 2000;283(6):749-55. PubMed PMID: 10683053.	Intervention does not include a diagnostic CPR
Matic I, Titlic M, Dikanovic M, Jurjevic M, Jukic I, Tonkic A. Effects of APACHE II score on mechanical ventilation; prediction and outcome. <i>Acta anaesthesiologica Belgica [Internet]</i> . 2007; 58(3):[177-83 pp.].	Intervention does not include a diagnostic CPR
Moylan CA, Brady CW, Johnson JL, Smith AD, Tuttle-Newhall JE, Muir AJ. Disparities in liver transplantation before and after introduction of the MELD score. <i>JAMA.</i> 2008 Nov 26;300(20):2371-8. PubMed PMID: 19033587. Pubmed Central PMCID: 3640479.	Intervention does not include a diagnostic CPR
Meyer G, Kopke S, Bender R, Muhlhauser I. Predicting the risk of falling--efficacy of a risk assessment tool compared to nurses' judgement: a cluster-randomised controlled trial [ISRCTN37794278]. <i>BMC Geriatr.</i> 2005;5:14. PubMed PMID: 16285880. Pubmed Central PMCID: 1312310.	Intervention does not include a diagnostic CPR

McGregor JC, Weekes E, Forrest GN, Standiford HC, Perencevich EN, Furuno JP, et al. Impact of a computerized clinical decision support system on reducing inappropriate antimicrobial use: a randomized controlled trial. Journal of the American Medical Informatics Association. 2006;13(4):378-84. PubMed PMID: 16622162. Pubmed Central PMCID: PMC1513678.	Intervention does not include a diagnostic CPR
Montgomery AA, Fahey T, Peters TJ, MacIntosh C, Sharp DJ. Evaluation of computer based clinical decision support system and risk chart for management of hypertension in primary care: randomised controlled trial. BMJ. 2000;320(7236):686-90. PubMed PMID: 10710578. Pubmed Central PMCID: PMC27312.	Intervention does not include a diagnostic CPR
Montgomery AA, Emmett CL, Fahey T, Jones C, Ricketts I, Patel RR, et al. Two decision aids for mode of delivery among women with previous caesarean section: randomised controlled trial. BMJ (Clinical research ed) [Internet]. 2007; 334(7607):[1305 p.].	Intervention does not include a diagnostic CPR
Murray LS, Teasdale GM, Murray GD, Jennett B, Miller JD, Pickard JD, et al. Does prediction of outcome alter patient management? Lancet. 1993 Jun 12;341(8859):1487-91. PubMed PMID: 8099377.	Intervention does not include a diagnostic CPR
Plank J, Blaha J, Cordingley J, Wilinska ME, Chassin LJ, Morgan C, et al. Multicentric, randomized, controlled trial to evaluate blood glucose control by the model predictive control algorithm versus routine glucose management protocols in intensive care unit patients. Diabetes care [Internet]. 2006; 29(2):[271-6 pp.].	Intervention does not include a diagnostic CPR
Paul M, Andreassen S, Tacconelli E, Nielsen AD, Almasreh N, Frank U, et al. Improving empirical antibiotic treatment using TREAT, a computerized decision support system: cluster randomized trial. J Antimicrob Chemother. 2006 Dec;58(6):1238-45. PubMed PMID: 16998208.	Intervention does not include a diagnostic CPR
Protheroe J, Bower P, Chew-Graham C, Peters TJ, Fahey T. Effectiveness of a computerized decision aid in primary care on decision making and quality of life in menorrhagia: results of the MENTIP randomized controlled trial. Medical decision making : an international journal of the Society for Medical Decision Making [Internet]. 2007; 27(5):[575-84 pp.].	Intervention does not include a diagnostic CPR
Radecki SE, Brunton SA. Randomized clinical trial of a diagnostic instrument for pain complaints. Family medicine [Internet]. 1999; 31(10):[713-21 pp.].	Intervention does not include a diagnostic CPR
Ross MA, Compton S, Medado P, Fitzgerald M, Kilanowski P, O'Neil BJ. An emergency department diagnostic protocol for patients with transient ischemic attack: a randomized controlled trial. Annals of emergency medicine [Internet]. 2007; 50(2):[109-19 pp.].	Intervention does not include a diagnostic CPR
Ross MA, Kilanowski P, Matke A, B ON, Compton S. The Emergency Department Transient Ischemic Attack Accelerated Diagnostic Protocol (TIA ADP) Study. Annals of Emergency Medicine [Internet]. 2004; 44(4 Suppl 1):[S121 p.].	Intervention does not include a diagnostic CPR
Schrager DL, Gibbons PS, Langone CA, Lee S, Altshuler LL. Enabling the diagnosis of occult psychiatric illness in the emergency department: a randomized, controlled trial of the computerized, self-administered PRIME-MD diagnostic system. Ann Emerg Med. 2001 Feb;37(2):132-40. PubMed PMID: 11174229.	Intervention does not include a diagnostic CPR as defined by review
Thomas RE, Croal BL, Ramsay C, Eccles M, Grimshaw J. Effect of enhanced feedback and brief educational reminder messages on laboratory test requesting in primary care: a cluster randomised trial. Lancet. 2006 Jun 17;367(9527):1990-6. PubMed PMID: 16782489.	Intervention does not include a diagnostic CPR
Visser FJ, van der Vegt MJ, van der Wilt GJ, Janssen JP. The optimization of the diagnostic work-up in patients with suspected obstructive lung disease. BMC Pulmonary Medicine. 2010;10:60. PubMed PMID: 21092293. Pubmed Central PMCID: PMC2996350.	Intervention does not include a diagnostic CPR
van Wijk MA, van der Lei J, Mosseveld M, Bohnen AM, van Bommel JH. Assessment of decision support for blood test ordering in primary care: a randomized trial. Ann Intern Med. 2001 Feb 20;134(4):274-81. PubMed PMID: 11182837.	Intervention does not include a diagnostic CPR
Study is a systematic review of 'decision support tools' (these were checked for relevant trials)	
Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Annals of Internal Medicine. 2012;157(1):29-43. PubMed PMID: 22751758.	References checked
Cleveringa FG, Gorter KJ, van den Donk M, van Gijssel J, Rutten GE. Computerized decision support systems in primary care for type 2 diabetes patients only improve patients' outcomes when combined with feedback on performance and case management: a systematic review. Diabetes Technology & Therapeutics. 2013;15(2):180-92. PubMed PMID: 23360424.	References checked
Fillmore CL, Bray BE, Kawamoto K. Systematic review of clinical decision support interventions with potential for inpatient cost reduction. BMC Med Inform Decis Mak. 2013;13:135. PubMed PMID: 24344752. Pubmed Central PMCID: 3878492.	References checked
Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005;293(10):1223-38. PubMed PMID: 15755945.	References checked
Hagiwara M, Henricson M, Jonsson A, Suserud BO. Decision-support tool in prehospital care: a systematic review of randomized trials. Prehospital & Disaster Medicine. 2011;26(5):319-29. PubMed PMID: 22030101.	References checked
Roshanov PS, Misra S, Gerstein HC, Garg AX, Sebaldt RJ, Mackay JA, et al. Computerized clinical decision support systems for chronic disease management: a decision-maker-researcher partnership systematic review. Implementation Science. 2011;6:92. PubMed PMID: 21824386. Pubmed Central PMCID: PMC3170626.	References checked
Roshanov PS, You JJ, Dhaliwal J, Koff D, Mackay JA, Weise-Kelly L, et al. Can computerized clinical decision support systems improve practitioners' diagnostic test ordering behavior? A decision-	References checked

maker-researcher partnership systematic review. Implementation Science. 2011;6:88. PubMed PMID: 21824382. Pubmed Central PMCID: PMC3174115.	
Stengel D, Bauwens K, Rademacher G, Ekkernkamp A, Guthoff C. Emergency ultrasound-based algorithms for diagnosing blunt abdominal trauma. The Cochrane database of systematic reviews. 2013;7:CD004446. PubMed PMID: 23904141.	References checked
Protocol or in-progress impact study	
Anderson RT, Montori VM, Shah ND, Ting HH, Pencille LJ, Demers M, et al. Effectiveness of the Chest Pain Choice decision aid in emergency department patients with low-risk chest pain: study protocol for a multicenter randomized trial. Trials. 2014;15:166. PubMed PMID: 24884807. Pubmed Central PMCID: 4031497.	Protocol
Hess EP, Wyatt KD, Kharbada AB, Louie JP, Dayan PS, Tzimenatos L, et al. Effectiveness of the head CT choice decision aid in parents of children with minor head trauma: study protocol for a multicenter randomized trial. Trials. 2014;15:253. PubMed PMID: 24965659. Pubmed Central PMCID: 4081461.	Protocol
Murray GD. Assessing the clinical impact of a predictive system in severe head injury. Medical informatics = Medecine et informatique. 1990 Jul-Sep;15(3):269-73. PubMed PMID: 2232962.	Protocol
Mann DM, Kannry JL, Edonyabo D, Li AC, Arciniega J, Stulman J, et al. Rationale, design, and implementation protocol of an electronic health record integrated clinical prediction rule (iCPR) randomized trial in primary care. Implementation science : IS [Internet]. 2011; 6:[109 p.].	Protocol for study of McGinn (included in review)
Pierce MA, Hess EP, Kline JA, Shah ND, Breslin M, Branda ME, et al. The Chest Pain Choice trial: a pilot randomized trial of a decision aid for patients with chest pain in the emergency department. Trials. 2010;11:57. PubMed PMID: 20478056. Pubmed Central PMCID: 2881067.	Protocol for study of Hess
Poldervaart JM, Reitsma JB, Koffijberg H, Backus BE, Six AJ, Doevendans PA, et al. The impact of the HEART risk score in the early assessment of patients with acute chest pain: design of a stepped wedge, cluster randomised trial. BMC Cardiovasc Disord. 2013;13:77. PubMed PMID: 24070098. Pubmed Central PMCID: 3849098.	In progress study
Stiell IG, Grimshaw J, Wells GA, Coyle D, Lesiuk HJ, Rowe BH, et al. A matched-pair cluster design study protocol to evaluate implementation of the Canadian C-spine rule in hospital emergency departments: Phase III. Implement Sci. 2007;2:4. PubMed PMID: 17288613. Pubmed Central PMCID: 1802999.	Protocol
Other	
Chusak O. Prediction of Late-Onset Neonatal Sepsis Using LNS Score Comparing with Physicians' Probability Estimates: A Cluster Randomized Trial. Pediatric Academic Society [Internet]. 2008; http://www.abstracts2view.com/pas/(469) .	Unable to obtain full text
Corey GA, Merenstein JH. Applying the acute ischemic heart disease predictive instrument. The Journal of family practice [Internet]. 1987; 25(2):[127-33 pp.].	Reference standard not current
Kurashima S, Kobayashi K, Toyabe S, Akazawa K. Accuracy and efficiency of computer-aided nursing diagnosis. International journal of nursing terminologies and classifications : the official journal of NANDA International [Internet]. 2008; 19(3):[95-101 pp.].	Study not in real patients
Murray GD, Murray LS, Barlow P, Teasdale GM, Jennett WB. Assessing the performance and clinical impact of a computerized prognostic system in severe head injury. Stat Med. 1986 Sep-Oct;5(5):403-10. PubMed PMID: 3538261.	Comparison in paper cases
Tierney WM, McDonald CJ, Hui SL, Martin DK. Computer predictions of abnormal test results. Effects on outpatient testing. JAMA. 1988 Feb 26;259(8):1194-8. PubMed PMID: 3339821.	CPR not for clinical diagnosis
Wexler JR, Swender PT, Tunnessen WW, Jr., Oski FA. Impact of a system of computer-assisted diagnosis. Initial evaluation of the hospitalized patient. Am J Dis Child. 1975 Feb;129(2):203-5. PubMed PMID: 1091140.	CPR for diagnosis across multiple body systems

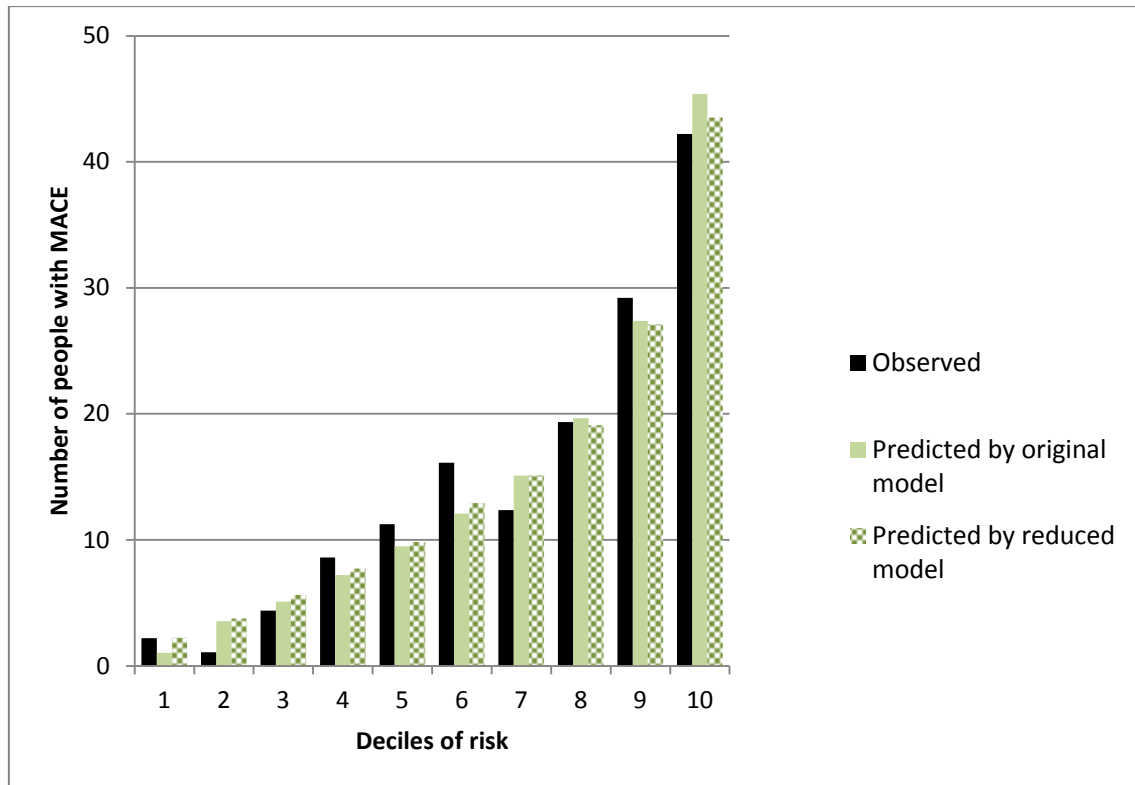
Types of primary and secondary outcomes reported in the included studies

Study	Patient outcomes	Process outcomes			
		Process of care	Clinicians Decisions	Accuracy	Use and application
Worrall et al 2007			X		
Mclsaac & Goel 1998			X		
Mclsaac et al 2002			X		
McGinn et al 2013			X		x
Little et al 2013	X		x		
Douglas et al 2000		X	x		
Farahnak et al 2007		X	x		
Lintula et al 2010				X	
Lintula et al 2009				X	
Wellwood et al 1992	x		x	X	
Roukema et al 2008		X			
Lacroix et al 2014		x	X		
de Vos-Kerkhof et al 2015		x	X		
Auleley et al 1997	x		X		x
Fan et al 2006	x	X	x		
Than et al 2014	x		X		
Sanchis et al 2010		X			
Torres et al 2014	x		X		
Klassen et al 1993		x	X		
Walter et al 2012		x	X	x	
Rodger et al 2006	X		x		
Horowitz et al 2007	X	x			
Bogusevicius et al 2002		x		X	
Stiell et al 2010			X	x	x
Stiell et al 2009			X		x

X primary outcome X secondary outcome

Appendix C Supplementary material from the study of the simplification of a clinical prediction rule

The observed and expected major adverse cardiac events for the original and reduced model by decile of predicted risk in the validation dataset



Calculation of reclassification indices

Unweighted score (cut-off ≥4)	
Event NRI	$\text{Pr}(\text{up event}) - \text{Pr}(\text{down event}) = (\text{number of events classified up} - \text{number of events classified down}) / \text{number of events}$ $(0 - 0) / 133 = 0$
Non-event NRI	$\text{Pr}(\text{down nonevent}) - \text{Pr}(\text{up nonevent}) = (\text{number of nonevents classified down} - \text{number of nonevents classified up}) / \text{number of nonevents}$ $(3 - 86) / 776 = -.10696 = 10.7\%$
Overall NRI	$[\text{Pr}(\text{up event}) - \text{Pr}(\text{down event})] + [\text{Pr}(\text{down nonevent}) - \text{Pr}(\text{up nonevent})] = \text{event NRI} + \text{nonevent NRI}$ $0 + -.10696 = -.10696 = -10.7\%$
Weighted NRI Based on decision threshold of 2% the relative weight of true positives to true negatives is 49:1. Therefore the corresponding weights for the event and non-event components of NRI become $s_1 = x + 1 = 50$ and $s_2 = (x + 1) / x = 1.02041$	$s_1 \times (\text{P}(\text{event up}) \times \text{P}(\text{up}) - \text{P}(\text{event down}) \times \text{P}(\text{down})) + s_2 \times (\text{P}(\text{nonevent down}) \times \text{P}(\text{down}) - \text{P}(\text{down}) - \text{P}(\text{nonevent up}) \times \text{P}(\text{up}))$ (Pencina 2011) Not calculated as there are no event reclassifications with the unweighted score
Reduced weighted score (cut-off ≥17)	
Event NRI	$\text{Pr}(\text{up event}) - \text{Pr}(\text{down event}) = (\text{number of events classified up} - \text{number of events classified down}) / \text{number of events}$

	$(1-0)/133 = .00752 = 0.752\%$
No-nevent NRI	$\text{Pr}(\text{downInnonevent}) - \text{Pr}(\text{upInnonevent}) = (\text{number of nonevents classified down} - \text{number of nonevents classified up})/\text{number of nonevents}$ $(24-98)/776 = -.0954 = -9.54\%$
Overall NRI	$[\text{Pr}(\text{upIevent}) - \text{Pr}(\text{downIevent})] + [\text{Pr}(\text{downInnonevent}) - \text{Pr}(\text{upInnonevent})] = \text{event NRI} + \text{nonevent NRI}$ $0.752 + -9.54 = -8.788$
Weighted NRI Based on decision threshold of 2% the relative weight of true positives to true negatives is 49:1. Therefore the corresponding weights for the event and non-event components of NRI become $s_1 = x + 1 = 50$ and $s_2 = (x + 1)/x = 1.02041$	$s_1 \times (\text{P}(\text{eventIup}) \times \text{P}(\text{up}) - \text{P}(\text{eventIdown}) \times \text{P}(\text{down})) + s_2 \times (\text{P}(\text{noneventIdown}) \times \text{P}(\text{down}) - \text{P}(\text{down}) - \text{P}(\text{noneventIup}) \times \text{P}(\text{up}))$ (Pencina 2011) $50 \times (132-131)/909 = .055005$ $+ 1.02041 \times (325-399)/909 = -.0830694$ $\text{wNRI} = .055005 + -.0830694 = -.028064 = -2.81$
Reduced unweighted score (cut-off ≥ 3)	
Event NRI	$\text{Pr}(\text{upIevent}) - \text{Pr}(\text{downIevent}) = (\text{number of events classified up} - \text{number of events classified down})/\text{number of events}$ $(2-0)/133 = .015038 = 1.504\%$
Non-event NRI	$\text{Pr}(\text{downInnonevent}) - \text{Pr}(\text{upInnonevent}) = (\text{number of nonevents classified down} - \text{number of nonevents classified up})/\text{number of nonevents}$ $(3-250)/776 = -.318299 = -31.8\%$
Overall NRI	$[\text{Pr}(\text{upIevent}) - \text{Pr}(\text{downIevent})] + [\text{Pr}(\text{downInnonevent}) - \text{Pr}(\text{upInnonevent})] = \text{event NRI} + \text{nonevent NRI}$ $1.504 + -31.8 = -30.29$
Weighted NRI Based on decision threshold of 2% the relative weight of true positives to true negatives is 49:1. Therefore the corresponding weights for the event and non-event components of NRI become $s_1 = x + 1 = 50$ and $s_2 = (x + 1)/x = 1.02041$	$s_1 \times (\text{P}(\text{eventIup}) \times \text{P}(\text{up}) - \text{P}(\text{eventIdown}) \times \text{P}(\text{down})) + s_2 \times (\text{P}(\text{noneventIdown}) \times \text{P}(\text{down}) - \text{P}(\text{down}) - \text{P}(\text{noneventIup}) \times \text{P}(\text{up}))$ (Pencina 2011) $50 \times (133-131)/909 = .11001100$ $+ 1.020408 \times (325-572)/909 = -.271727172$ $\text{wNRI} = .11001100 + -.271727172 = -.1617161709 = -16.172$

Reference list

1. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine*. 2015 Jan 6;162(1):W1-73. PubMed PMID: 25560730.
2. Steyerberg EW. *Clinical prediction models. A practical approach to development, validation and updating*. New York: Springer; 2009.
3. Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. *Journal of the American Medical Informatics Association : JAMIA*. 2007 Mar-Apr;14(2):141-5. PubMed PMID: 17213487. Pubmed Central PMCID: 2213467.
4. Stacey D, Legare F, Col NF, Bennett CL, Barry MJ, Eden KB, et al. Decision aids for people facing health treatment or screening decisions. *The Cochrane database of systematic reviews*. 2014;1:CD001431. PubMed PMID: 24470076.
5. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of internal medicine*. 2006 Feb 7;144(3):201-9. PubMed PMID: 16461965.
6. Steyerberg EW, Van Calster B, Pencina MJ. [Performance measures for prediction models and markers: evaluation of predictions and classifications]. *Revista espanola de cardiologia*. 2011 Sep;64(9):788-94. PubMed PMID: 21763052. Medidas del rendimiento de modelos de prediccion y marcadores pronosticos: evaluacion de las predicciones y clasificaciones.
7. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. *Annals of emergency medicine*. 1992 Apr;21(4):384-90. PubMed PMID: 1554175.
8. McIsaac WJ, White D, Tannenbaum D, Low DE. A clinical score to reduce unnecessary antibiotic use in patients with sore throat. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 1998 Jan 13;158(1):75-83. PubMed PMID: 9475915. Pubmed Central PMCID: 1228750.
9. Haskins R, Osmotherly PG, Rivett DA. Diagnostic clinical prediction rules for specific subtypes of low back pain: a systematic review. *The Journal of orthopaedic and sports physical therapy*. 2015 Feb;45(2):61-76, A1-4. PubMed PMID: 25573009.
10. Hendriksen JM, Geersing GJ, Moons KG, de Groot JA. Diagnostic and prognostic prediction models. *Journal of thrombosis and haemostasis : JTH*. 2013 Jun;11 Suppl 1:129-41. PubMed PMID: 23809117.
11. Ebell M. AHRQ White Paper: Use of clinical decision rules for point-of-care decision support. *Medical decision making : an international journal of the Society for Medical Decision Making*. 2010 Nov-Dec;30(6):712-21. PubMed PMID: 21183758.
12. Broekhuizen BD, Sachs A, Janssen K, Geersing GJ, Moons K, Hoes A, et al. Does a decision aid help physicians to detect chronic obstructive pulmonary disease? *The British journal of general practice : the journal of the Royal College of General Practitioners*. 2011 Oct;61(591):e674-9. PubMed PMID: 22152850. Pubmed Central PMCID: 3177137.
13. Croskerry P. A universal model of diagnostic reasoning. *Academic medicine : journal of the Association of American Medical Colleges*. 2009 Aug;84(8):1022-8. PubMed PMID: 19638766.

14. Balla J, Heneghan C, Thompson M, Balla M. Clinical decision making in a high-risk primary care environment: a qualitative study in the UK. *BMJ open*. 2012;2:e000414. PubMed PMID: 22318661. Pubmed Central PMCID: 3330259.
15. Balla JI, Heneghan C, Glasziou P, Thompson M, Balla ME. A model for reflection for good clinical practice. *Journal of evaluation in clinical practice*. 2009 Dec;15(6):964-9. PubMed PMID: 20367693.
16. Croskerry P. Critical thinking and reasoning in emergency medicine. In: Croskerry P, Cosby KS, Schenkel SM, Wears RL, editors. *Patient safety in emergency medicine*. Philadelphia (PA): Lippincott Williams & Wilkins; 2008.
17. Pelaccia T, Tardif J, Tribby E, Charlin B. An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. *Medical education online*. 2011;16. PubMed PMID: 21430797. Pubmed Central PMCID: 3060310.
18. Graber ML. The incidence of diagnostic error in medicine. *BMJ quality & safety*. 2013 Oct;22 Suppl 2:ii21-ii7. PubMed PMID: 23771902. Pubmed Central PMCID: 3786666.
19. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Archives of internal medicine*. 2005 Jul 11;165(13):1493-9. PubMed PMID: 16009864.
20. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *The American journal of medicine*. 2008 May;121(5 Suppl):S2-23. PubMed PMID: 18440350.
21. Singh H, Giardina TD, Meyer AN, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. *JAMA internal medicine*. 2013 Mar 25;173(6):418-25. PubMed PMID: 23440149. Pubmed Central PMCID: 3690001.
22. Saber Tehrani AS, Lee H, Mathews SC, Shore A, Makary MA, Pronovost PJ, et al. 25-Year summary of US malpractice claims for diagnostic errors 1986-2010: an analysis from the National Practitioner Data Bank. *BMJ quality & safety*. 2013 Aug;22(8):672-80. PubMed PMID: 23610443.
23. Singh H, Weingart SN. Diagnostic errors in ambulatory care: dimensions and preventive strategies. *Advances in health sciences education : theory and practice*. 2009 Sep;14 Suppl 1:57-61. PubMed PMID: 19669923. Pubmed Central PMCID: 3643195.
24. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine : journal of the Association of American Medical Colleges*. 2003 Aug;78(8):775-80. PubMed PMID: 12915363.
25. Croskerry P, Singhal G, Mamede S. Cognitive debiasing 1: origins of bias and theory of debiasing. *BMJ quality & safety*. 2013 Oct;22 Suppl 2:ii58-ii64. PubMed PMID: 23882089. Pubmed Central PMCID: 3786658.
26. Norman GR, Eva KW. Diagnostic error and clinical reasoning. *Medical education*. 2010 Jan;44(1):94-100. PubMed PMID: 20078760.
27. Blumenthal-Barby JS, Krieger H. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. *Medical decision making : an international journal of the Society for Medical Decision Making*. 2015 May;35(4):539-57. PubMed PMID: 25145577.
28. Bornstein BH, Emler AC. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of evaluation in clinical practice*. 2001 May;7(2):97-107. PubMed PMID: 11489035.

29. Zwaan L, Schiff GD, Singh H. Advancing the research agenda for diagnostic error reduction. *BMJ quality & safety*. 2013 Oct;22 Suppl 2:ii52-ii7. PubMed PMID: 23942182. Pubmed Central PMCID: 3786655.
30. Minue S, Bermudez-Tamayo C, Fernandez A, Martin-Martin JJ, Benitez V, Melguizo M, et al. Identification of factors associated with diagnostic error in primary care. *BMC family practice*. 2014;15:92. PubMed PMID: 24884984. Pubmed Central PMCID: 4024115.
31. Zwaan L, Thijs A, Wagner C, van der Wal G, Timmermans DR. Relating faults in diagnostic reasoning with diagnostic errors and patient harm. *Academic medicine : journal of the Association of American Medical Colleges*. 2012 Feb;87(2):149-56. PubMed PMID: 22189886.
32. Moons KG, de Groot JA, Linnet K, Reitsma JB, Bossuyt PM. Quantifying the added value of a diagnostic test or marker. *Clinical chemistry*. 2012 Oct;58(10):1408-17. PubMed PMID: 22952348.
33. Friedman MH, Connell KJ, Olthoff AJ, Sinacore JM, Bordage G. Medical student errors in making a diagnosis. *Academic medicine : journal of the Association of American Medical Colleges*. 1998 Oct;73(10 Suppl):S19-21. PubMed PMID: 9795640.
34. Gruppen LD, Wolf FM, Billi JE. Information gathering and integration as sources of error in diagnostic decision making. *Medical decision making : an international journal of the Society for Medical Decision Making*. 1991 Oct-Dec;11(4):233-9. PubMed PMID: 1766327.
35. Richardson WS, Wilson MC. Textbook descriptions of disease--where's the beef? *ACP journal club*. 2002 Jul-Aug;137(1):A11-2. PubMed PMID: 12093237.
36. Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. *Science*. 1989 Mar 31;243(4899):1668-74. PubMed PMID: 2648573.
37. Henriksen K, Brady J. The pursuit of better diagnostic performance: a human factors perspective. *BMJ quality & safety*. 2013 Oct;22 Suppl 2:ii1-ii5. PubMed PMID: 23704082. Pubmed Central PMCID: 3786636.
38. Newman-Toker DE, Pronovost PJ. Diagnostic errors--the next frontier for patient safety. *Jama*. 2009 Mar 11;301(10):1060-2. PubMed PMID: 19278949.
39. Graber ML, Kissam S, Payne VL, Meyer AN, Sorensen A, Lenfestey N, et al. Cognitive interventions to reduce diagnostic error: a narrative review. *BMJ quality & safety*. 2012 Jul;21(7):535-57. PubMed PMID: 22543420.
40. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. *The New England journal of medicine*. 1985 Sep 26;313(13):793-9. PubMed PMID: 3897864.
41. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. *Jama*. 2000 Jul 5;284(1):79-84. PubMed PMID: 10872017.
42. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. 2012 May;98(9):683-90. PubMed PMID: 22397945.
43. Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R, Hedges T. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. *AMIA Annual Symposium proceedings / AMIA Symposium*. 2003:728-32. PubMed PMID: 14728269. Pubmed Central PMCID: 1479983.

44. Keogh C, Wallace E, O'Brien KK, Galvin R, Smith SM, Lewis C, et al. Developing an international register of clinical prediction rules for use in primary care: a descriptive analysis. *Annals of family medicine*. 2014 Jul;12(4):359-66. PubMed PMID: 25024245. Pubmed Central PMCID: 4096474.
45. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS medicine*. 2013;10(2):e1001381. PubMed PMID: 23393430. Pubmed Central PMCID: 3564751.
46. Pluddemann A, Wallace E, Bankhead C, Keogh C, Van der Windt D, Lasserson D, et al. Clinical prediction rules in practice: review of clinical guidelines and survey of GPs. *The British journal of general practice : the journal of the Royal College of General Practitioners*. 2014 Apr;64(621):e233-42. PubMed PMID: 24686888. Pubmed Central PMCID: 3964449.
47. Authors/Task Force m, Elliott PM, Anastasakis A, Borger MA, Borggrefe M, Cecchi F, et al. 2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy: the Task Force for the Diagnosis and Management of Hypertrophic Cardiomyopathy of the European Society of Cardiology (ESC). *European heart journal*. 2014 Oct 14;35(39):2733-79. PubMed PMID: 25173338.
48. Rabar S, Lau R, O'Flynn N, Li L, Barry P, Guideline Development G. Risk assessment of fragility fractures: summary of NICE guidance. *Bmj*. 2012;345:e3698. PubMed PMID: 22875946.
49. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. *Health technology assessment*. 2010 Feb;14(8):iii, ix-xi, 1-193. PubMed PMID: 20181324.
50. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *Bmj*. 2012;344:e686. PubMed PMID: 22354600.
51. Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. *Medical decision making : an international journal of the Society for Medical Decision Making*. 2009 Sep-Oct;29(5):E30-8. PubMed PMID: 19726782.
52. Fahey T, van der Lei J. Producing and using clinical prediction rules. In: Knottnerus JA, Buntinx F, editors. *The Evidence Base of Clinical Diagnosis: Theory and Methods of Diagnostic Research*. 2nd ed: Blackwell Publishing Ltd.; 2009.
53. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *The New England journal of medicine*. 1980 May 15;302(20):1109-17. PubMed PMID: 7366635.
54. Reilly BM. Physical examination in the care of medical inpatients: an observational study. *Lancet*. 2003 Oct 4;362(9390):1100-5. PubMed PMID: 14550696.
55. Aegisdóttir S, White M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R.S., Nichols, C.N., Lamropoulos, G.K., Walker, B.S., Cohen, G., Rush, J.D. The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*. 2006;34:341-82.
56. Gigerenzer G, Todd, P.M., & the ABC Research Group. *Simple heuristics that make us smart*. New York: Oxford University Press; 1999.
57. Grove WM, Meehl PE. Comparative efficiency of informal (subjective, impressionistic), and formal (mechanical, algorithmic) prediction procedures; The clinical-statistical controversy. *Psychology, Public Policy and Law*. 1996;2(293-323).
58. Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*. 2000 Mar;12(1):19-30. PubMed PMID: 10752360.

59. Holt RR. Clinical and statistical prediction: A reformulation and some new data. *The journal of abnormal and social psychology*. 1958;56:1-12.
60. Sanchis J, Bosch X, Bodi V, Nunez J, Doltra A, Heras M, et al. Randomized comparison between clinical evaluation plus N-terminal pro-B-type natriuretic peptide versus exercise testing for decision making in acute chest pain of uncertain origin. *American heart journal*. 2010 Feb;159(2):176-82. PubMed PMID: 20152214.
61. Bossuyt PMM, McCaffery K. Additional Patient Outcomes and Pathways in Evaluations of Testing. *Medical Tests-White Paper Series. AHRQ Methods for Effective Health Care*. Rockville (MD)2009.
62. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012 May;98(9):691-8. PubMed PMID: 22397946.
63. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal*. 2014 Aug 1;35(29):1925-31. PubMed PMID: 24898551. Pubmed Central PMCID: 4155437.
64. Wallace E, Smith SM, Perera-Salazar R, Vaucher P, McCowan C, Collins G, et al. Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs). *BMC medical informatics and decision making*. 2011;11:62. PubMed PMID: 21999201. Pubmed Central PMCID: 3216240.
65. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. *Bmj*. 2002 Feb 23;324(7335):477-80. PubMed PMID: 11859054. Pubmed Central PMCID: 1122397.
66. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Bmj*. 2009;338:b606. PubMed PMID: 19502216.
67. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS medicine*. 2014 Oct;11(10):e1001744. PubMed PMID: 25314315. Pubmed Central PMCID: 4196729.
68. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in medicine*. 2007 Jul 10;26(15):2937-57. PubMed PMID: 17186501.
69. Koon S, Petscher Y. Comparing methodologies for developing an early warning system: Classification and regression tree model versus logistic regression (REL 2015-077). Washington DC. : U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.; 2015.
70. Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB. A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. *Journal of investigative medicine : the official publication of the American Federation for Clinical Research*. 1995 Oct;43(5):468-76. PubMed PMID: 8528758.
71. Tsien CL, Fraser HS, Long WJ, Kennedy RL. Using classification tree and logistic regression methods to diagnose myocardial infarction. *Studies in health technology and informatics*. 1998;52 Pt 1:493-7. PubMed PMID: 10384505.
72. Knuiman MW, Vu HT, Segal MR. An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. *Journal of cardiovascular risk*. 1997 Apr;4(2):127-34. PubMed PMID: 9304494.

73. Cole T. Scaling and rounding regression-coefficients to integers. *Applied statistics*. 1993;42(1):261-68.
74. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of clinical epidemiology*. 2015 Mar;68(3):279-89. PubMed PMID: 25179855.
75. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. *Statistical methods in medical research*. 2014 Apr 7. PubMed PMID: 23907781. Pubmed Central PMCID: 3933449.
76. Pencina MJ, D'Agostino RB, Sr. Evaluating Discrimination of Risk Prediction Models: The C Statistic. *Jama*. 2015 Sep 8;314(10):1063-4. PubMed PMID: 26348755.
77. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clinical chemistry*. 2008 Jan;54(1):17-23. PubMed PMID: 18024533.
78. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2013 Aug 6;185(11):E537-44. PubMed PMID: 23798453. Pubmed Central PMCID: 3735771.
79. Kline JA, Johnson CL, Pollack CV, Jr., Diercks DB, Hollander JE, Newgard CD, et al. Pretest probability assessment derived from attribute matching. *BMC medical informatics and decision making*. 2005;5:26. PubMed PMID: 16095534. Pubmed Central PMCID: 1201143.
80. Ebell MH, Locatelli I, Senn N. A novel approach to the determination of clinical decision thresholds. *Evidence-based medicine*. 2015 Apr;20(2):41-7. PubMed PMID: 25736042.
81. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical decision making : an international journal of the Society for Medical Decision Making*. 2006 Nov-Dec;26(6):565-74. PubMed PMID: 17099194. Pubmed Central PMCID: 2577036.
82. Leening MJ, Vedder MM, Witteman JC, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Annals of internal medicine*. 2014 Jan 21;160(2):122-31. PubMed PMID: 24592497.
83. Muhlenbruch K, Heraclides A, Steyerberg EW, Joost HG, Boeing H, Schulze MB. Assessing improvement in disease prediction using net reclassification improvement: impact of risk cut-offs and number of risk categories. *European journal of epidemiology*. 2013 Jan;28(1):25-33. PubMed PMID: 23179629.
84. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW. Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. *Medical decision making : an international journal of the Society for Medical Decision Making*. 2013 May;33(4):490-501. PubMed PMID: 23313931. Pubmed Central PMCID: 4066820.
85. Linnet K, Bossuyt PM, Moons KG, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. *Clinical chemistry*. 2012 Sep;58(9):1292-301. PubMed PMID: 22829313.
86. Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. *Journal of clinical epidemiology*. 2010 Aug;63(8):883-91. PubMed PMID: 20079607.
87. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *Bmj*. 2006 May 6;332(7549):1089-92. PubMed PMID: 16675820. Pubmed Central PMCID: 1458557.

88. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC medical research methodology*. 2011;11:13. PubMed PMID: 21276237. Pubmed Central PMCID: 3042425.
89. Kappen TH, Vergouwe Y, van Wolfswinkel L, Kalkman CJ, Moons KG, van Klei WA. Impact of adding therapeutic recommendations to risk assessments from a prediction model for postoperative nausea and vomiting. *British journal of anaesthesia*. 2015 Feb;114(2):252-60. PubMed PMID: 25274048.
90. Graham ID, Stiell IG, Laupacis A, McAuley L, Howell M, Clancy M, et al. Awareness and use of the Ottawa ankle and knee rules in 5 countries: can publication alone be enough to change practice? *Annals of emergency medicine*. 2001 Mar;37(3):259-66. PubMed PMID: 11223761.
91. Koffijberg H, van Zaane B, Moons KG. From accuracy to patient outcome and cost-effectiveness evaluations of diagnostic tests and biomarkers: an exemplary modelling study. *BMC medical research methodology*. 2013;13:12. PubMed PMID: 23368927. Pubmed Central PMCID: 3724486.
92. Hunink MG. Decision making in the face of uncertainty and resource constraints: examples from trauma imaging. *Radiology*. 2005 May;235(2):375-83. PubMed PMID: 15858081.
93. Siontis KC, Siontis GC, Contopoulos-Ioannidis DG, Ioannidis JP. Patient outcomes in randomized comparisons of diagnostic tests: still the ultimate judge: Reply to letter by Ferrante di Ruffano and Deeks "Test-treatment RCTs - sheep in wolves' clothing". *Journal of clinical epidemiology*. 2015 Jun 27. PubMed PMID: 26130596.
94. Penaloza A, Verschuren F, Meyer G, Quentin-Georget S, Soulie C, Thys F, et al. Comparison of the unstructured clinician gestalt, the wells score, and the revised Geneva score to estimate pretest probability for suspected pulmonary embolism. *Annals of emergency medicine*. 2013 Aug;62(2):117-24 e2. PubMed PMID: 23433653.
95. Haskins R, Osmotherly PG, Southgate E, Rivett DA. Physiotherapists' knowledge, attitudes and practices regarding clinical prediction rules for low back pain. *Manual therapy*. 2014 Apr;19(2):142-51. PubMed PMID: 24176916.
96. Brehaut JC, Stiell IG, Graham ID. Will a new clinical decision rule be widely used? The case of the Canadian C-spine rule. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2006 Apr;13(4):413-20. PubMed PMID: 16531607.
97. Arkes HR, Shaffer VA, Medow MA. The influence of a physician's use of a diagnostic decision aid on the malpractice verdicts of mock jurors. *Medical decision making : an international journal of the Society for Medical Decision Making*. 2008 Mar-Apr;28(2):201-8. PubMed PMID: 18349437.
98. Arkes HR, Shaffer VA, Medow MA. Patients derogate physicians who use a computer-assisted diagnostic aid. *Medical decision making : an international journal of the Society for Medical Decision Making*. 2007 Mar-Apr;27(2):189-202. PubMed PMID: 17409368.
99. Kleinrouweler CE, Cheong-See FM, Collins GS, Kwee A, Thangaratinam S, Khan KS, et al. Prognostic models in obstetrics: available, but far from applicable. *American journal of obstetrics and gynecology*. 2015 Jun 10. PubMed PMID: 26070707.
100. Maguire JL, Kulik DM, Laupacis A, Kuppermann N, Uleryk EM, Parkin PC. Clinical prediction rules for children: a systematic review. *Pediatrics*. 2011 Sep;128(3):e666-77. PubMed PMID: 21859912.
101. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *Jama*. 1997 Feb 12;277(6):488-94. PubMed PMID: 9020274.

102. Vickers AJ, Cronin AM. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). *Urology*. 2010 Dec;76(6):1298-301. PubMed PMID: 21030068. Pubmed Central PMCID: 2997853.
103. Collins G. Introducing the Tripod Statement for Reporting Clinical Prediction Models 2015 Jan 22 [cited 22 January 2015]. In: blogs.plos.org [internet]. Plos Medical Journals' Community Blog. 2015. Available from: blogs.plos.org/speakingofmedicine/2015/01/22/introducing-tripod-statement-reporting-clinical-prediction-models/.
104. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS medicine*. 2012;9(5):1-12. PubMed PMID: 22629234. Pubmed Central PMCID: 3358324.
105. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC medical research methodology*. 2014;14:40. PubMed PMID: 24645774. Pubmed Central PMCID: 3999945.
106. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine*. 2011;9:103. PubMed PMID: 21902820. Pubmed Central PMCID: 3180398.
107. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. *BMC medicine*. 2010;8:20. PubMed PMID: 20353578. Pubmed Central PMCID: 2856521.
108. Wyatt JC, Altman DG. Commentary: Prognostic models: clinically useful or quickly forgotten? *Bmj*. 1995;311(7019):1539-41.
109. Brehaut JC, Stiell IG, Visentin L, Graham ID. Clinical decision rules "in the real world": how a widely disseminated rule is used in everyday practice. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2005 Oct;12(10):948-56. PubMed PMID: 16166599.
110. Costantino MM, Randall G, Gosselin M, Brandt M, Spinning K, Vegas CD. CT angiography in the evaluation of acute pulmonary embolus. *AJR American journal of roentgenology*. 2008 Aug;191(2):471-4. PubMed PMID: 18647919.
111. Heneghan C, Glasziou P, Thompson M, Rose P, Balla J, Lasserson D, et al. Diagnostic strategies used in primary care. *Bmj*. 2009;338:b946. PubMed PMID: 19380414. Pubmed Central PMCID: 3266845.
112. Eagles D, Stiell IG, Clement CM, Brehaut J, Taljaard M, Kelly AM, et al. International survey of emergency physicians' awareness and use of the Canadian Cervical-Spine Rule and the Canadian Computed Tomography Head Rule. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2008 Dec;15(12):1256-61. PubMed PMID: 18945241.
113. Dowling SK, Wishart I. Use of the Ottawa Ankle Rules in children: a survey of physicians' practice patterns. *Cjem*. 2011 Sep;13(5):333-8; E44-6. PubMed PMID: 21955415.
114. Graham ID, Stiell IG, Laupacis A, O'Connor AM, Wells GA. Emergency physicians' attitudes toward and use of clinical decision rules for radiography. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 1998 Feb;5(2):134-40. PubMed PMID: 9492134.
115. Runyon MS, Richman PB, Kline JA, Pulmonary Embolism Research Consortium Study G. Emergency medicine practitioner knowledge and use of decision rules for the evaluation of patients with suspected pulmonary embolism: variations by practice setting and training level. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2007 Jan;14(1):53-7. PubMed PMID: 17119186.

116. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PA, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. *Jama*. 1999 Oct 20;282(15):1458-65. PubMed PMID: 10535437.
117. Woolf SH. Practice guidelines: a new reality in medicine. III. Impact on patient care. *Archives of internal medicine*. 1993 Dec 13;153(23):2646-55. PubMed PMID: 8250661.
118. Boutis K, Constantine E, Schuh S, Pecaric M, Stephens D, Narayanan UG. Pediatric emergency physician opinions on ankle radiograph clinical decision rules. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2010 Jul;17(7):709-17. PubMed PMID: 20653584.
119. Bero LA, Grilli R, Grimshaw JM, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. *Bmj*. 1998 Aug 15;317(7156):465-8. PubMed PMID: 9703533. Pubmed Central PMCID: 1113716.
120. Rabin BA, Glasgow RE, Kerner JF, Klump MP, Brownson RC. Dissemination and implementation research on community-based cancer prevention: a systematic review. *American journal of preventive medicine*. 2010 Apr;38(4):443-56. PubMed PMID: 20307814.
121. Cohen JF, Cohen R, Levy C, Thollot F, Benani M, Bidet P, et al. Selective testing strategies for diagnosing group A streptococcal infection in children with pharyngitis: a systematic review and prospective multicentre external validation study. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2015 Jan 6;187(1):23-32. PubMed PMID: 25487666. Pubmed Central PMCID: 4284164.
122. Lucassen W, Geersing GJ, Erkens PM, Reitsma JB, Moons KG, Buller H, et al. Clinical decision rules for excluding pulmonary embolism: a meta-analysis. *Annals of internal medicine*. 2011 Oct 4;155(7):448-60. PubMed PMID: 21969343.
123. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. *Bmj*. 2011;343:d7163. PubMed PMID: 22123912. Pubmed Central PMCID: 3225074.
124. Collins GS, Moons KG. Comparing risk prediction models. *Bmj*. 2012;344:e3186. PubMed PMID: 22628131.
125. Keogh C, Fahey T. Clinical prediction rules in primary care: what can be done to maximise their implementation? *Clinical Evidence [Internet]*. 2010.
126. Ballard DW, Rauchwerger AS, Reed ME, Vinson DR, Mark DG, Offerman SR, et al. Emergency physicians' knowledge and attitudes of clinical decision support in the electronic health record: a survey-based study. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2013 Apr;20(4):352-60. PubMed PMID: 23701342.
127. Pearson SD, Goldman L, Garcia TB, Cook EF, Lee TH. Physician response to a prediction rule for the triage of emergency department patients with chest pain. *Journal of general internal medicine*. 1994 May;9(5):241-7. PubMed PMID: 8046525.
128. Eichler K, Zoller M, Tschudi P, Steurer J. Barriers to apply cardiovascular prediction rules in primary care: a postal survey. *BMC family practice*. 2007;8:1. PubMed PMID: 17201905. Pubmed Central PMCID: 1766351.
129. Beutel BG, Trehan SK, Shalvoy RM, Mello MJ. The Ottawa knee rule: examining use in an academic emergency department. *The western journal of emergency medicine*. 2012 Sep;13(4):366-72. PubMed PMID: 23251717. Pubmed Central PMCID: 3523897.

130. Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. *Annals of emergency medicine*. 1999 Apr;33(4):437-47. PubMed PMID: 10092723.
131. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *Journal of clinical epidemiology*. 2008 Nov;61(11):1085-94. PubMed PMID: 19208371.
132. Haskins R, Osmotherly PG, Southgate E, Rivett DA. Australian physiotherapists' priorities for the development of clinical prediction rules for low back pain: a qualitative study. *Physiotherapy*. 2015 Mar;101(1):44-9. PubMed PMID: 25037535.
133. Curran JA, Brehaut J, Patey AM, Osmond M, Stiell I, Grimshaw JM. Understanding the Canadian adult CT head rule trial: use of the theoretical domains framework for process evaluation. *Implementation science* : IS. 2013;8:25. PubMed PMID: 23433082. Pubmed Central PMCID: 3585785.
134. Drescher FS, Chandrika S, Weir ID, Weintraub JT, Berman L, Lee R, et al. Effectiveness and acceptability of a computerized decision support system using modified Wells criteria for evaluation of suspected pulmonary embolism. *Annals of emergency medicine*. 2011 Jun;57(6):613-21. PubMed PMID: 21050624.
135. Bonner C, Jansen J, McKinn S, Irwig L, Doust J, Glasziou P, et al. General practitioners' use of different cardiovascular risk assessment strategies: a qualitative study. *The Medical journal of Australia*. 2013 Oct 7;199(7):485-9. PubMed PMID: 24099210.
136. Eagles D, Stiell IG, Clement CM, Brehaut J, Kelly AM, Mason S, et al. International survey of emergency physicians' priorities for clinical decision rules. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2008 Feb;15(2):177-82. PubMed PMID: 18275448.
137. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*. 2011 Oct 18;155(8):529-36. PubMed PMID: 22007046.
138. Al Omar MZ, Baldwin GA. Reappraisal of use of X-rays in childhood ankle and midfoot injuries. *Emergency radiology*. 2002 Jul;9(2):88-92. PubMed PMID: 15290584.
139. Attia MW, Zaoutis T, Klein JD, Meier FA. Performance of a predictive model for streptococcal pharyngitis in children. *Archives of pediatrics & adolescent medicine*. 2001 Jun;155(6):687-91. PubMed PMID: 11386959.
140. Bigaroni A, Perrier A, Bounameaux H. Is clinical probability assessment of deep vein thrombosis by a score really standardized? *Thrombosis and haemostasis*. 2000 May;83(5):788-9. PubMed PMID: 10823281.
141. Blattler W, Martinez I, Blattler IK. Diagnosis of deep venous thrombosis and alternative diseases in symptomatic outpatients. *European journal of internal medicine*. 2004 Aug;15(5):305-11. PubMed PMID: 15450988.
142. Bojang KA, Obaro S, Morison LA, Greenwood BM. A prospective evaluation of a clinical algorithm for the diagnosis of malaria in Gambian children. *Tropical medicine & international health : TM & IH*. 2000 Apr;5(4):231-6. PubMed PMID: 10810013.
143. Carrier M, Wells PS, Rodger MA. Excluding pulmonary embolism at the bedside with low pre-test probability and D-dimer: safety and clinical utility of 4 methods to assign pre-test probability. *Thrombosis research*. 2006;117(4):469-74. PubMed PMID: 15893807.

144. Cebul RD, Poses RM. The comparative cost-effectiveness of statistical decision rules and experienced physicians in pharyngitis management. *Jama*. 1986 Dec 26;256(24):3353-7. PubMed PMID: 3097339.
145. Chagnon I, Bounameaux H, Aujesky D, Roy PM, Gourdier AL, Cornuz J, et al. Comparison of two clinical prediction rules and implicit assessment among patients with suspected pulmonary embolism. *The American journal of medicine*. 2002 Sep;113(4):269-75. PubMed PMID: 12361811.
146. Cornuz J, Ghali WA, Hayoz D, Stoianov R, Depairon M, Yersin B. Clinical prediction of deep venous thrombosis using two risk assessment methods in combination with rapid quantitative D-dimer testing. *The American journal of medicine*. 2002 Feb 15;112(3):198-203. PubMed PMID: 11893346.
147. Crowe L, Anderson V, Babl FE. Application of the CHALICE clinical prediction rule for intracranial injury in children outside the UK: impact on head CT rate. *Archives of disease in childhood*. 2010 Dec;95(12):1017-22. PubMed PMID: 20573733.
148. El-Solh AA, Hsiao CB, Goodnough S, Serghani J, Grant BJ. Predicting active pulmonary tuberculosis using an artificial neural network. *Chest*. 1999 Oct;116(4):968-73. PubMed PMID: 10531161.
149. Emerman CL, Dawson N, Speroff T, Siciliano C, Effron D, Rashad F, et al. Comparison of physician judgment and decision aids for ordering chest radiographs for pneumonia in outpatients. *Annals of emergency medicine*. 1991 Nov;20(11):1215-9. PubMed PMID: 1952308.
150. Fenyo G. Routine use of a scoring system for decision-making in suspected acute appendicitis in adults. *Acta chirurgica Scandinavica*. 1987 Sep;153(9):545-51. PubMed PMID: 3321809.
151. Geersing GJ, Janssen KJ, Oudega R, van Weert H, Stoffers H, Hoes A, et al. Diagnostic classification in patients with suspected deep venous thrombosis: physicians' judgement or a decision rule? *The British journal of general practice : the journal of the Royal College of General Practitioners*. 2010 Oct;60(579):742-8. PubMed PMID: 20883623. Pubmed Central PMCID: 2944933.
152. Glas AS, Pijnenburg BA, Lijmer JG, Bogaard K, de RM, Keeman JN, et al. Comparison of diagnostic decision rules and structured data collection in assessment of acute ankle injury. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2002 Mar 19;166(6):727-33. PubMed PMID: 11944759. Pubmed Central PMCID: 99451.
153. Kabrhel C, Mark Courtney D, Camargo CA, Jr., Moore CL, Richman PB, Plewa MC, et al. Potential impact of adjusting the threshold of the quantitative D-dimer based on pretest probability of acute pulmonary embolism. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2009 Apr;16(4):325-32. PubMed PMID: 19298619.
154. Kabrhel C, McAfee AT, Goldhaber SZ. The contribution of the subjective component of the Canadian Pulmonary Embolism Score to the overall score in emergency department patients. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2005 Oct;12(10):915-20. PubMed PMID: 16204134.
155. Kline JA, Courtney DM, Kabrhel C, Moore CL, Smithline HA, Plewa MC, et al. Prospective multicenter evaluation of the pulmonary embolism rule-out criteria. *Journal of thrombosis and haemostasis : JTH*. 2008 May;6(5):772-80. PubMed PMID: 18318689.
156. Leibovici L, Greenshtain S, Cohen O, Mor F, Wysenbeek AJ. Bacteremia in febrile patients. A clinical model for diagnosis. *Archives of internal medicine*. 1991 Sep;151(9):1801-6. PubMed PMID: 1888246.

157. Meltzer AC, Baumann BM, Chen EH, Shofer FS, Mills AM. Poor sensitivity of a modified Alvarado score in adults with suspected appendicitis. *Annals of emergency medicine*. 2013 Aug;62(2):126-31. PubMed PMID: 23623557.
158. Miron MJ, Perrier A, Bounameaux H. Clinical assessment of suspected deep vein thrombosis: comparison between a score and empirical assessment. *Journal of internal medicine*. 2000 Feb;247(2):249-54. PubMed PMID: 10692088.
159. Mitchell AM, Garvey JL, Chandra A, Diercks D, Pollack CV, Kline JA. Prospective multicenter study of quantitative pretest probability assessment to exclude acute coronary syndrome for patients evaluated in emergency department chest pain units. *Annals of emergency medicine*. 2006 May;47(5):447. PubMed PMID: 16631984.
160. Penalzoa A, Verschuren F, Dambrine S, Zech F, Thys F, Roy PM. Performance of the Pulmonary Embolism Rule-out Criteria (the PERC rule) combined with low clinical probability in high prevalence population. *Thrombosis research*. 2012 May;129(5):e189-93. PubMed PMID: 22424852.
161. Rosenberg P, McIsaac W, Macintosh D, Kroll M. Diagnosing streptococcal pharyngitis in the emergency department: Is a sore throat score approach better than rapid streptococcal antigen testing? *Cjem*. 2002 May;4(3):178-84. PubMed PMID: 17609003.
162. Runyon MS, Webb WB, Jones AE, Kline JA. Comparison of the unstructured clinician estimate of pretest probability for pulmonary embolism to the Canadian score and the Charlotte rule: a prospective observational study. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2005 Jul;12(7):587-93. PubMed PMID: 15995088.
163. Sanson BJ, Lijmer JG, Mac Gillavry MR, Turkstra F, Prins MH, Buller HR. Comparison of a clinical probability estimate and two clinical models in patients with suspected pulmonary embolism. ANTELOPE-Study Group. *Thrombosis and haemostasis*. 2000 Feb;83(2):199-203. PubMed PMID: 10739372.
164. Singh-Ranger G, Marathias A. Comparison of current local practice and the Ottawa Ankle Rules to determine the need for radiography in acute ankle injury. *Accident and emergency nursing*. 1999 Oct;7(4):201-6. PubMed PMID: 10808759.
165. Stein J, Louie J, Flanders S, Maselli J, Hacker JK, Drew WL, et al. Performance characteristics of clinical diagnosis, a clinical decision rule, and a rapid influenza test in the detection of influenza infection in a community sample of adults. *Annals of emergency medicine*. 2005 Nov;46(5):412-9. PubMed PMID: 16271670.
166. Vaillancourt C, Stiell IG, Beaudoin T, Maloney J, Anton AR, Bradford P, et al. The out-of-hospital validation of the Canadian C-Spine Rule by paramedics. *Annals of emergency medicine*. 2009 Nov;54(5):663-71 e1. PubMed PMID: 19394111.
167. Wang B, Lin Y, Pan FS, Yao C, Zheng ZY, Cai D, et al. Comparison of empirical estimate of clinical pretest probability with the Wells score for diagnosis of deep vein thrombosis. *Blood coagulation & fibrinolysis : an international journal in haemostasis and thrombosis*. 2013 Jan;24(1):76-81. PubMed PMID: 23103729.
168. Korner H, Sondenaa K, Soreide JA, Andersen E, Nysted A, Lende TH. Structured data collection improves the diagnosis of acute appendicitis. *The British journal of surgery*. 1998 Mar;85(3):341-4. PubMed PMID: 9529488.
169. Kabrhel C, Camargo CA, Jr., Goldhaber SZ. Clinical gestalt and the diagnosis of pulmonary embolism: does experience matter? *Chest*. 2005 May;127(5):1627-30. PubMed PMID: 15888838.

170. Liu JL, Wyatt JC, Deeks JJ, Clamp S, Keen J, Verde P, et al. Systematic reviews of clinical decision tools for acute abdominal pain. *Health technology assessment*. 2006 Nov;10(47):1-167, iii-iv. PubMed PMID: 17083855.
171. Auleley GR, Ravaud P, Giraudeau B, Kerboull L, Nizard R, Massin P, et al. Implementation of the Ottawa ankle rules in France. A multicenter randomized controlled trial. *Jama*. 1997 Jun 25;277(24):1935-9. PubMed PMID: 9200633.
172. Siontis KC, Siontis GC, Contopoulos-Ioannidis DG, Ioannidis JP. Diagnostic tests often fail to lead to changes in patient outcomes. *Journal of clinical epidemiology*. 2014 Jun;67(6):612-21. PubMed PMID: 24679598.
173. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. *Annals of internal medicine*. 2012 Jul 3;157(1):29-43. PubMed PMID: 22751758.
174. Holstiege J, Mathes T, Pieper D. Effects of computer-aided clinical decision support systems in improving antibiotic prescribing by primary care providers: a systematic review. *Journal of the American Medical Informatics Association : JAMIA*. 2015 Jan;22(1):236-42. PubMed PMID: 25125688.
175. Roshanov PS, You JJ, Dhaliwal J, Koff D, Mackay JA, Weise-Kelly L, et al. Can computerized clinical decision support systems improve practitioners' diagnostic test ordering behavior? A decision-maker-researcher partnership systematic review. *Implementation science : IS*. 2011;6:88. PubMed PMID: 21824382. Pubmed Central PMCID: 3174115.
176. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama*. 2005 Mar 9;293(10):1223-38. PubMed PMID: 15755945.
177. Jaspers MW, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *Journal of the American Medical Informatics Association : JAMIA*. 2011 May 1;18(3):327-34. PubMed PMID: 21422100. Pubmed Central PMCID: 3078663.
178. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P, Group C. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Annals of internal medicine*. 2008 Feb 19;148(4):295-309. PubMed PMID: 18283207.
179. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*. 2010;8:18. PubMed PMID: 20334633. Pubmed Central PMCID: 2860339.
180. Higgins JPT, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.1 [Updated March 2011]*: The Cochrane Collaboration; 2011.
181. Higgins JPT, (editors) GS. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]*: The Cochrane Collaboration; 2011.
182. Little P, Hobbs FD, Moore M, Mant D, Williamson I, McNulty C, et al. Clinical score and rapid antigen detection test to guide antibiotic use for sore throats: randomised controlled trial of PRISM (primary care streptococcal management). *Bmj*. 2013;347:f5806. PubMed PMID: 24114306. Pubmed Central PMCID: 3805475.
183. Worrall G, Hutchinson J, Sherman G, Griffiths J. Diagnosing streptococcal sore throat in adults: randomized controlled trial of in-office aids. *Canadian family physician Medecin de famille canadien*. 2007 Apr;53(4):666-71. PubMed PMID: 17872717. Pubmed Central PMCID: 1952596.

184. Wellwood J, Johannessen S, Spiegelhalter DJ. How does computer-aided diagnosis improve the management of acute abdominal pain? *Annals of the Royal College of Surgeons of England*. 1992 Jan;74(1):40-6. PubMed PMID: 1736794. Pubmed Central PMCID: 2497469.
185. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clinical trials*. 2005;2(2):99-107. PubMed PMID: 16279131.
186. Bogusevicius A, Maleckas A, Pundzius J, Skaudickas D. Prospective randomised trial of computer-aided diagnosis and contrast radiography in acute small bowel obstruction. *The European journal of surgery = Acta chirurgica*. 2002;168(2):78-83. PubMed PMID: 12113275.
187. Douglas CD, Macpherson NE, Davidson PM, Gani JS. Randomised controlled trial of ultrasonography in diagnosis of acute appendicitis, incorporating the Alvarado score. *Bmj*. 2000 Oct 14;321(7266):919-22. PubMed PMID: 11030676. Pubmed Central PMCID: 27498.
188. Fan J, Woolfrey K. The effect of triage-applied Ottawa Ankle Rules on the length of stay in a Canadian urgent care department: a randomized controlled trial. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2006 Feb;13(2):153-7. PubMed PMID: 16436790.
189. Farahnak M, Talaei-Khoei M, Gorouhi F, Jalali A, Gorouhi F. The Alvarado score and antibiotics therapy as a corporate protocol versus conventional clinical management: randomized controlled pilot study of approach to acute appendicitis. *The American journal of emergency medicine*. 2007 Sep;25(7):850-2. PubMed PMID: 17870498.
190. Horowitz N, Moshkowitz M, Leshno M, Ribak J, Birkenfeld S, Kenet G, et al. Clinical trial: evaluation of a clinical decision-support model for upper abdominal complaints in primary-care practice. *Alimentary pharmacology & therapeutics*. 2007 Nov 1;26(9):1277-83. PubMed PMID: 17944742.
191. Klassen TP, Ropp LJ, Sutcliffe T, Blouin R, Dulberg C, Raman S, et al. A randomized, controlled trial of radiograph ordering for extremity trauma in a pediatric emergency department. *Annals of emergency medicine*. 1993 Oct;22(10):1524-9. PubMed PMID: 8214829.
192. Lintula H, Kokki H, Kettunen R, Eskelinen M. Appendicitis score for children with suspected appendicitis. A randomized clinical trial. *Langenbeck's archives of surgery / Deutsche Gesellschaft fur Chirurgie*. 2009 Nov;394(6):999-1004. PubMed PMID: 18841382.
193. Lintula H, Kokki H, Pulkkinen J, Kettunen R, Grohn O, Eskelinen M. Diagnostic score in acute appendicitis. Validation of a diagnostic score (Lintula score) for adults with suspected appendicitis. *Langenbeck's archives of surgery / Deutsche Gesellschaft fur Chirurgie*. 2010 Jun;395(5):495-500. PubMed PMID: 20379739.
194. McGinn TG, McCullagh L, Kannry J, Knaus M, Sofianou A, Wisnivesky JP, et al. Efficacy of an evidence-based clinical decision support in primary care practices: a randomized clinical trial. *JAMA internal medicine*. 2013 Sep 23;173(17):1584-91. PubMed PMID: 23896675.
195. McIsaac WJ, Goel V. Effect of an explicit decision-support tool on decisions to prescribe antibiotics for sore throat. *Medical decision making : an international journal of the Society for Medical Decision Making*. 1998 Apr-Jun;18(2):220-8. PubMed PMID: 9566455.
196. McIsaac WJ, Goel V, To T, Permaul JA, Low DE. Effect on antibiotic prescribing of repeated clinical prompts to use a sore throat score: lessons from a failed community intervention study. *The Journal of family practice*. 2002 Apr;51(4):339-44. PubMed PMID: 11978257.
197. Rodger MA, Bredeson CN, Jones G, Rasuli P, Raymond F, Clement AM, et al. The bedside investigation of pulmonary embolism diagnosis study: a double-blind randomized controlled trial

comparing combinations of 3 bedside tests vs ventilation-perfusion scan for the initial investigation of suspected pulmonary embolism. *Archives of internal medicine*. 2006 Jan 23;166(2):181-7. PubMed PMID: 16432086.

198. Roukema J, Steyerberg EW, van der Lei J, Moll HA. Randomized trial of a clinical decision support system: impact on the management of children with fever without apparent source. *Journal of the American Medical Informatics Association : JAMIA*. 2008 Jan-Feb;15(1):107-13. PubMed PMID: 17947627. Pubmed Central PMCID: 2273109.

199. Stiell IG, Clement CM, Grimshaw J, Brison RJ, Rowe BH, Schull MJ, et al. Implementation of the Canadian C-Spine Rule: prospective 12 centre cluster randomised trial. *Bmj*. 2009;339:b4146. PubMed PMID: 19875425. Pubmed Central PMCID: 2770593.

200. Stiell IG, Clement CM, Grimshaw JM, Brison RJ, Rowe BH, Lee JS, et al. A prospective cluster-randomized trial to implement the Canadian CT Head Rule in emergency departments. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2010 Oct 5;182(14):1527-32. PubMed PMID: 20732978. Pubmed Central PMCID: 2950184.

201. Than M, Aldous S, Lord SJ, Goodacre S, Frampton CM, Troughton R, et al. A 2-hour diagnostic protocol for possible cardiac chest pain in the emergency department: a randomized clinical trial. *JAMA internal medicine*. 2014 Jan;174(1):51-8. PubMed PMID: 24100783.

202. Torres FA, Pasarelli I, Cutri A, Ossorio MF, Ferrero F. Impact assessment of a decision rule for using antibiotics in pneumonia: a randomized trial. *Pediatric pulmonology*. 2014 Jul;49(7):701-6. PubMed PMID: 24039234.

203. Walter FM, Morris HC, Humphrys E, Hall PN, Prevost AT, Burrows N, et al. Effect of adding a diagnostic aid to best practice to manage suspicious pigmented lesions in primary care: randomised controlled trial. *Bmj*. 2012;345:e4110. PubMed PMID: 22763392. Pubmed Central PMCID: 3389518.

204. de Vos-Kerkhof E, Nijman RG, Vergouwe Y, Polinder S, Steyerberg EW, van der Lei J, et al. Impact of a clinical decision model for febrile children at risk for serious bacterial infections at the emergency department: a randomized controlled trial. *PloS one*. 2015;10(5):e0127620. PubMed PMID: 26024532. Pubmed Central PMCID: 4449197.

205. Lacroix L, Manzano S, Vandertuin L, Hugon F, Galetto-Lacour A, Gervais A. Impact of the lab-score on antibiotic prescription rate in children with fever without source: a randomized controlled trial. *PloS one*. 2014;9(12):e115061. PubMed PMID: 25503770. Pubmed Central PMCID: 4263728.

206. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*. 2014 Jan 11;383(9912):166-75. PubMed PMID: 24411645. Pubmed Central PMCID: 4697939.

207. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JP. Reproducible Research Practices and Transparency across the Biomedical Literature. *PLoS biology*. 2016 Jan;14(1):e1002333. PubMed PMID: 26726926. Pubmed Central PMCID: 4699702.

208. Deeks JJ. Assessing outcomes following tests. In: Price CP, Christenson RH, editors. *Evidence-based laboratory medicine: principles, practice and outcomes*. 2nd ed. Washington DC: AACC Press; 2007. p. 95-111.

209. Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. *Bmj*. 2012;345:e5661. PubMed PMID: 22951546.

210. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Bmj*. 2010;340:c869. PubMed PMID: 20332511. Pubmed Central PMCID: 2844943.

211. Bryant J, Passey ME, Hall AE, Sanson-Fisher RW. A systematic review of the quality of reporting in published smoking cessation trials for pregnant women: an explanation for the evidence-practice gap? *Implementation science* : IS. 2014;9:94. PubMed PMID: 25138616. Pubmed Central PMCID: 4147164.
212. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? *Bmj*. 2008 Jun 28;336(7659):1472-4. PubMed PMID: 18583680. Pubmed Central PMCID: 2440840.
213. Delaney A, Angus DC, Bellomo R, Cameron P, Cooper DJ, Finfer S, et al. Bench-to-bedside review: the evaluation of complex interventions in critical care. *Critical care*. 2008;12(2):210. PubMed PMID: 18439321. Pubmed Central PMCID: 2447586.
214. Smelt AF, van der Weele GM, Blom JW, Gussekloo J, Assendelft WJ. How usual is usual care in pragmatic intervention studies in primary care? An overview of recent trials. *The British journal of general practice : the journal of the Royal College of General Practitioners*. 2010 Jul;60(576):e305-18. PubMed PMID: 20594432. Pubmed Central PMCID: 2894405.
215. Hay AD, Heron J, Ness A, team As. The prevalence of symptoms and consultations in pre-school children in the Avon Longitudinal Study of Parents and Children (ALSPAC): a prospective cohort study. *Family practice*. 2005 Aug;22(4):367-74. PubMed PMID: 15897210.
216. Alpern ER, Stanley RM, Gorelick MH, Donaldson A, Knight S, Teach SJ, et al. Epidemiology of a pediatric emergency medicine research network: the PECARN Core Data Project. *Pediatric emergency care*. 2006 Oct;22(10):689-99. PubMed PMID: 17047467.
217. Armon K, Stephenson T, Gabriel V, MacFaul R, Eccleston P, Werneke U, et al. Determining the common medical presenting problems to an accident and emergency department. *Archives of disease in childhood*. 2001 May;84(5):390-2. PubMed PMID: 11316679. Pubmed Central PMCID: 1718762.
218. Massin MM, Montesanti J, Gerard P, Lepage P. Spectrum and frequency of illness presenting to a pediatric emergency department. *Acta clinica Belgica*. 2006 Jul-Aug;61(4):161-5. PubMed PMID: 17091911.
219. Van den Bruel A, Bartholomeeusen S, Aertgeerts B, Truyers C, Buntinx F. Serious infections in children: an incidence study in family practice. *BMC family practice*. 2006;7:23. PubMed PMID: 16569232. Pubmed Central PMCID: 1435901.
220. Craig JC, Williams GJ, Jones M, Codarini M, Macaskill P, Hayen A, et al. The accuracy of clinical symptoms and signs for the diagnosis of serious bacterial infection in young febrile children: prospective cohort study of 15 781 febrile illnesses. *Bmj*. 2010;340:c1594. PubMed PMID: 20406860. Pubmed Central PMCID: 2857748.
221. Thompson MJ, Ninis N, Perera R, Mayon-White R, Phillips C, Bailey L, et al. Clinical recognition of meningococcal disease in children and adolescents. *Lancet*. 2006 Feb 4;367(9508):397-403. PubMed PMID: 16458763.
222. Thompson M, Van den Bruel A, Verbakel J, Lakhanpaul M, Haj-Hassan T, Stevens R, et al. Systematic review and validation of prediction rules for identifying children with serious infections in emergency departments and urgent-access primary care. *Health technology assessment*. 2012;16(15):1-100. PubMed PMID: 22452986.
223. Van den Bruel A, Haj-Hassan T, Thompson M, Buntinx F, Mant D, European Research Network on Recognising Serious Infection i. Diagnostic value of clinical features at presentation to identify serious infection in children in developed countries: a systematic review. *Lancet*. 2010 Mar 6;375(9717):834-45. PubMed PMID: 20132979.

224. Elshout G, van Ierland Y, Bohnen AM, de Wilde M, Moll HA, Oostenbrink R, et al. Alarming signs and symptoms in febrile children in primary care: an observational cohort study in The Netherlands. *PloS one*. 2014;9(2):e88114. PubMed PMID: 24586305. Pubmed Central PMCID: 3929539.
225. Oostenbrink R, Thompson M, Steyerberg EW, members E. Barriers to translating diagnostic research in febrile children to clinical practice: a systematic review. *Archives of disease in childhood*. 2012 Jul;97(7):667-72. PubMed PMID: 22219168.
226. Buntinx F, Mant D, Van den Bruel A, Donner-Banzhof N, Dinant GJ. Dealing with low-incidence serious diseases in general practice. *The British journal of general practice : the journal of the Royal College of General Practitioners*. 2011 Jan;61(582):43-6. PubMed PMID: 21401991. Pubmed Central PMCID: 3020049.
227. Coenen S, Van Royen P, Vermeire E, Hermann I, Denekens J. Antibiotics for coughing in general practice: a qualitative decision analysis. *Family practice*. 2000 Oct;17(5):380-5. PubMed PMID: 11021895.
228. Truyers C, Goderis G, Dewitte H, Akker M, Buntinx F. The Intego database: background, methods and basic results of a Flemish general practice-based continuous morbidity registration project. *BMC medical informatics and decision making*. 2014;14:48. PubMed PMID: 24906941. Pubmed Central PMCID: 4067630.
229. Costelloe C, Metcalfe C, Lovering A, Mant D, Hay AD. Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. *Bmj*. 2010;340:c2096. PubMed PMID: 20483949.
230. Goossens H, Ferech M, Vander Stichele R, Elseviers M, Group EP. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet*. 2005 Feb 12-18;365(9459):579-87. PubMed PMID: 15708101.
231. Librizzi J, McCulloh R, Koehn K, Alverson B. Appropriateness of testing for serious bacterial infection in children hospitalized with bronchiolitis. *Hospital pediatrics*. 2014 Jan;4(1):33-8. PubMed PMID: 24435599.
232. Reeves G. C-reactive protein. *Australian Prescriber*. 2007;30:74-6.
233. van Vugt SF, Broekhuizen BD, Lammens C, Zuithoff NP, de Jong PA, Coenen S, et al. Use of serum C reactive protein and procalcitonin concentrations in addition to symptoms and signs to predict pneumonia in patients presenting to primary care with acute cough: diagnostic study. *Bmj*. 2013;346:f2450. PubMed PMID: 23633005. Pubmed Central PMCID: 3639712.
234. Melbye H, Hvidsten D, Holm A, Nordbo SA, Brox J. The course of C-reactive protein response in untreated upper respiratory tract infection. *The British journal of general practice : the journal of the Royal College of General Practitioners*. 2004 Sep;54(506):653-8. PubMed PMID: 15353049. Pubmed Central PMCID: 1326064.
235. Melbye S, Stocks N. Point-of-care testing for C-reactive protein - a new path for Australian GPs? *Australian Family Physician*. 2006;35(7):513-17.
236. Bleeker SE, Moons KG, Derksen-Lubsen G, Grobbee DE, Moll HA. Predicting serious bacterial infection in young children with fever without apparent source. *Acta paediatrica*. 2001 Nov;90(11):1226-32. PubMed PMID: 11808890.
237. McCarthy PL, Sharpe MR, Spiesel SZ, Dolan TF, Forsyth BW, DeWitt TG, et al. Observation scales to identify serious illness in febrile children. *Pediatrics*. 1982 Nov;70(5):802-9. PubMed PMID: 7133831.

238. Crain EF, Bulas D, Bijur PE, Goldman HS. Is a chest radiograph necessary in the evaluation of every febrile infant less than 8 weeks of age? *Pediatrics*. 1991 Oct;88(4):821-4. PubMed PMID: 1896292.
239. Joffe A, McCormick M, DeAngelis C. Which children with febrile seizures need lumbar puncture? A decision analysis approach. *American journal of diseases of children*. 1983 Dec;137(12):1153-6. PubMed PMID: 6637930.
240. Offringa M, Beishuizen A, Derksen-Lubsen G, Lubsen J. Seizures and fever: can we rule out meningitis on clinical grounds alone? *Clinical pediatrics*. 1992 Sep;31(9):514-22. PubMed PMID: 1468167.
241. Oostenbrink R, Moons KG, Donders AR, Grobbee DE, Moll HA. Prediction of bacterial meningitis in children with meningeal signs: reduction of lumbar punctures. *Acta paediatrica*. 2001 Jun;90(6):611-7. PubMed PMID: 11440091.
242. Pantell RH, Newman TB, Bernzweig J, Bergman DA, Takayama JJ, Segal M, et al. Management and outcomes of care of fever in early infancy. *Jama*. 2004 Mar 10;291(10):1203-12. PubMed PMID: 15010441.
243. Schwartz RH, Wientzen RL, Jr. Occult bacteremia in toxic-appearing, febrile infants. A prospective clinical study in an office setting. *Clinical pediatrics*. 1982 Nov;21(11):659-63. PubMed PMID: 7127988.
244. Lacour AG, Gervaix A, Zamora SA, Vadas L, Lombard PR, Dayer JM, et al. Procalcitonin, IL-6, IL-8, IL-1 receptor antagonist and C-reactive protein as identifiers of serious bacterial infections in children with fever without localising signs. *European journal of pediatrics*. 2001 Feb;160(2):95-100. PubMed PMID: 11271398.
245. Galetto-Lacour A, Zamora SA, Gervaix A. Bedside procalcitonin and C-reactive protein tests in children with fever without localizing signs of infection seen in a referral center. *Pediatrics*. 2003 Nov;112(5):1054-60. PubMed PMID: 14595045.
246. Thayyil S, Shenoy M, Hamaluba M, Gupta A, Frater J, Verber IG. Is procalcitonin useful in early diagnosis of serious bacterial infections in children? *Acta paediatrica*. 2005 Feb;94(2):155-8. PubMed PMID: 15981747.
247. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *Journal of clinical epidemiology*. 2008 Jan;61(1):76-86. PubMed PMID: 18083464.
248. Janssen KJ, Donders AR, Harrell FE, Jr., Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: to impute is better than to ignore. *Journal of clinical epidemiology*. 2010 Jul;63(7):721-7. PubMed PMID: 20338724.
249. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology*. 2006 Oct;59(10):1087-91. PubMed PMID: 16980149.
250. van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of clinical epidemiology*. 2006 Oct;59(10):1102-9. PubMed PMID: 16980151.
251. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*. 2009;338:b2393. PubMed PMID: 19564179. Pubmed Central PMCID: 2714692.

252. Mason S, Bryan E, Bircher T, Clesham K, Eagles D, Clement C, et al. UK survey of emergency physicians' priorities for clinical decision rules. *Journal of Paramedic Practice*. 2011;3(2):78-82.
253. National Institute for Health and Clinical Excellence. Feverish illness: assessment and initial management in children younger than 5 years. NICE; 2007.
254. Nijman RG, Vergouwe Y, Thompson M, van Veen M, van Meurs AH, van der Lei J, et al. Clinical prediction model to aid emergency doctors managing febrile children at risk of serious bacterial infections: diagnostic study. *Bmj*. 2013;346:f1706. PubMed PMID: 23550046. Pubmed Central PMCID: 3614186.
255. Van den Bruel A, Aertgeerts B, Bruyninckx R, Aerts M, Buntinx F. Signs and symptoms for diagnosis of serious infections in children: a prospective study in primary care. *The British journal of general practice : the journal of the Royal College of General Practitioners*. 2007 Jul;57(540):538-46. PubMed PMID: 17727746. Pubmed Central PMCID: 2099636.
256. Bleeker SE, Derksen-Lubsen G, Grobbee DE, Donders AR, Moons KG, Moll HA. Validating and updating a prediction rule for serious bacterial infection in patients with fever without source. *Acta paediatrica*. 2007 Jan;96(1):100-4. PubMed PMID: 17187613.
257. Verbakel JY, Lemiengre MB, De Burghgraeve T, De Sutter A, Aertgeerts B, Bullens DM, et al. Validating a decision tree for serious infection: diagnostic accuracy in acutely ill children in ambulatory care. *BMJ open*. 2015;5(8):e008657. PubMed PMID: 26254472. Pubmed Central PMCID: 4538259.
258. Oostenbrink R, Thompson M, Lakhanpaul M, Steyerberg EW, Coad N, Moll HA. Children with fever and cough at emergency care: diagnostic accuracy of a clinical model to identify children at low risk of pneumonia. *European journal of emergency medicine : official journal of the European Society for Emergency Medicine*. 2013 Aug;20(4):273-80. PubMed PMID: 22868746.
259. King C. Evaluation and management of febrile infants in the emergency department. *Emergency medicine clinics of North America*. 2003 Feb;21(1):89-99, vi-vii. PubMed PMID: 12630733.
260. Ciesla G, Leader S, Stoddard J. Antibiotic prescribing rates in the US ambulatory care setting for patients diagnosed with influenza, 1997-2001. *Respiratory medicine*. 2004 Nov;98(11):1093-101. PubMed PMID: 15526810.
261. Gaur AH, Hare ME, Shorr RI. Provider and practice characteristics associated with antibiotic use in children with presumed viral respiratory tract infections. *Pediatrics*. 2005 Mar;115(3):635-41. PubMed PMID: 15741365.
262. Nyquist AC, Gonzales R, Steiner JF, Sande MA. Antibiotic prescribing for children with colds, upper respiratory tract infections, and bronchitis. *Jama*. 1998 Mar 18;279(11):875-7. PubMed PMID: 9516004.
263. Lindback S, Hellgren U, Julander I, Hansson LO. The value of C-reactive protein as a marker of bacterial infection in patients with septicaemia/endocarditis and influenza. *Scandinavian journal of infectious diseases*. 1989;21(5):543-9. PubMed PMID: 2587955.
264. Unkila-Kallio L, Kallio MJ, Eskola J, Peltola H. Serum C-reactive protein, erythrocyte sedimentation rate, and white blood cell count in acute hematogenous osteomyelitis of children. *Pediatrics*. 1994 Jan;93(1):59-62. PubMed PMID: 8265325.
265. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Annals of internal medicine*. 2004 Feb 3;140(3):189-202. PubMed PMID: 14757617.

266. Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health technology assessment*. 2004 Jun;8(25):iii, 1-234. PubMed PMID: 15193208.
267. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*. 2007 Apr;8(2):239-51. PubMed PMID: 16698768.
268. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology*. 2003 Nov;56(11):1129-35. PubMed PMID: 14615004.
269. Cobben JM, Cornelissen PJ, Haverkorn M, Waelkens JJ. [CRP versus BSE in pediatrics. How good is a diagnostic test?]. *Tijdschrift voor kindergeneeskunde*. 1990 Oct;58(5):169-74. PubMed PMID: 2247880. CRP versus BSE in de kindergeneeskunde. Hoe goed is een diagnostische test?
270. Isaacman DJ, Burke BL. Utility of the serum C-reactive protein for detection of occult bacterial infection in children. *Archives of pediatrics & adolescent medicine*. 2002 Sep;156(9):905-9. PubMed PMID: 12197798.
271. McCarthy PL, Jekel JF, Dolan TF, Jr. Comparison of acute-phase reactants in pediatric patients with fever. *Pediatrics*. 1978 Nov;62(5):716-20. PubMed PMID: 724316.
272. Pulliam PN, Attia MW, Cronan KM. C-reactive protein in febrile children 1 to 36 months of age with clinically undetectable serious bacterial infection. *Pediatrics*. 2001 Dec;108(6):1275-9. PubMed PMID: 11731648.
273. Andreola B, Bressan S, Callegaro S, Liverani A, Plebani M, Da Dalt L. Procalcitonin and C-reactive protein as diagnostic markers of severe bacterial infections in febrile infants and children in the emergency department. *The Pediatric infectious disease journal*. 2007 Aug;26(8):672-7. PubMed PMID: 17848876.
274. Berger RM, Berger MY, van Steensel-Moll HA, Dzoljic-Danilovic G, Derksen-Lubsen G. A predictive model to estimate the risk of serious bacterial infections in febrile infants. *European journal of pediatrics*. 1996 Jun;155(6):468-73. PubMed PMID: 8789763.
275. Tejani NR, Chonmaitree T, Rassin DK, Howie VM, Owen MJ, Goldman AS. Use of C-reactive protein in differentiation between acute bacterial and viral otitis media. *Pediatrics*. 1995 May;95(5):664-9. PubMed PMID: 7724300.
276. Fernandez Lopez A, Luaces Cubells C, Garcia Garcia JJ, Fernandez Pou J, Spanish Society of Pediatric E. Procalcitonin in pediatric emergency departments for the early diagnosis of invasive bacterial infections in febrile infants: results of a multicenter study and utility of a rapid qualitative test for this marker. *The Pediatric infectious disease journal*. 2003 Oct;22(10):895-903. PubMed PMID: 14551491.
277. Korppi M, Koskela M, Jalonen E, Leinonen M. Serologically indicated pneumococcal respiratory infection in children. *Scandinavian journal of infectious diseases*. 1992;24(4):437-43. PubMed PMID: 1411309.
278. Pratt A, Attia MW. Duration of fever and markers of serious bacterial infection in young febrile children. *Pediatrics international : official journal of the Japan Pediatric Society*. 2007 Feb;49(1):31-5. PubMed PMID: 17250502.
279. Chenillot O, Henny J, Steinmetz J, Herbeth B, Wagner C, Siest G. High sensitivity C-reactive protein: biological variations and reference limits. *Clinical chemistry and laboratory medicine : CCLM / FESCC*. 2000 Oct;38(10):1003-11. PubMed PMID: 11140615.

280. Foy R, Warner P. About time: diagnostic guidelines that help clinicians. Quality & safety in health care. 2003 Jun;12(3):205-9. PubMed PMID: 12792011. Pubmed Central PMCID: 1743712.
281. Codarini M, Craig J, Jones M. P, Williams G, Group. FS. Comparative performance of clinical judgment versus a diagnostic algorithm for serious bacterial infection (SBI) in children with fever. Pediatric Academic Societies' Annual Meeting; Toronto, Canada. May 5-8, 2007.
282. Herrera P, Duffau G. [Usefulness of C-reactive protein for the diagnosis of bacterial infections in children. A review]. Revista medica de Chile. 2005 May;133(5):541-6. PubMed PMID: 15970978. Existen bases para el uso de la proteina C reactiva en la deteccion de infecciones bacterianas en ninos?
283. van der Meer V, Neven AK, van den Broek PJ, Assendelft WJ. Diagnostic value of C reactive protein in infections of the lower respiratory tract: systematic review. Bmj. 2005 Jul 2;331(7507):26. PubMed PMID: 15979984. Pubmed Central PMCID: 558535.
284. Nijman RG, Moll HA, Vergouwe Y, de Rijke YB, Oostenbrink R. C-Reactive Protein Bedside Testing in Febrile Children Lowers Length of Stay at the Emergency Department. Pediatric emergency care. 2015 Sep;31(9):633-9. PubMed PMID: 26181498.
285. Aabenhus R, Jensen JU, Jorgensen KJ, Hrobjartsson A, Bjerrum L. Biomarkers as point-of-care tests to guide prescription of antibiotics in patients with acute respiratory infections in primary care. The Cochrane database of systematic reviews. 2014;11:CD010130. PubMed PMID: 25374293.
286. Verbakel JY, Lemiengre MB, De Burghgraeve T, De Sutter A, Bullens DM, Aertgeerts B, et al. Diagnosing serious infections in acutely ill children in ambulatory care (ERNIE 2 study protocol, part A): diagnostic accuracy of a clinical decision tree and added value of a point-of-care C-reactive protein test and oxygen saturation. BMC pediatrics. 2014;14:207. PubMed PMID: 25277457. Pubmed Central PMCID: 4287386.
287. Lemiengre MB, Verbakel JY, De Burghgraeve T, Aertgeerts B, De Baets F, Buntinx F, et al. Optimizing antibiotic prescribing for acutely ill children in primary care (ERNIE2 study protocol, part B): a cluster randomized, factorial controlled trial evaluating the effect of a point-of-care C-reactive protein test and a brief intervention combined with written safety net advice. BMC pediatrics. 2014;14:246. PubMed PMID: 25277543. Pubmed Central PMCID: 4287591.
288. Staub LP, Dyer S, Lord SJ, Simes RJ. Linking the evidence: intermediate outcomes in medical test assessments. International journal of technology assessment in health care. 2012 Jan;28(1):52-8. PubMed PMID: 22230006.
289. Antman EM, Cohen M, Bernink PJ, McCabe CH, Horacek T, Papuchis G, et al. The TIMI risk score for unstable angina/non-ST elevation MI: A method for prognostication and therapeutic decision making. Jama. 2000 Aug 16;284(7):835-42. PubMed PMID: 10938172.
290. TIMI Study Group. TIMI Risk Score Boston: TIMI Study Group; 2011 [17/08/2014].
291. Xavier Scheuermeyer F, Wong H, Yu E, Boychuk B, Innes G, Grafstein E, et al. Development and validation of a prediction rule for early discharge of low-risk emergency department patients with potential ischemic chest pain. Cjem. 2014 Mar 1;16(2):106-19. PubMed PMID: 24626115.
292. Fine AM, Brownstein JS, Nigrovic LE, Kimia AA, Olson KL, Thompson AD, et al. Integrating spatial epidemiology into a decision model for evaluation of facial palsy in children. Archives of pediatrics & adolescent medicine. 2011 Jan;165(1):61-7. PubMed PMID: 21199982. Pubmed Central PMCID: 3644029.
293. Zuithoff NP, Vergouwe Y, King M, Nazareth I, Hak E, Moons KG, et al. A clinical prediction rule for detecting major depressive disorder in primary care: the PREDICT-NL study. Family practice. 2009 Aug;26(4):241-50. PubMed PMID: 19546117.

294. Nijdam ME, Janssen KJ, Moons KG, Grobbee DE, van der Post JA, Bots ML, et al. Prediction model for hypertension in pregnancy in nulliparous women using information obtained at the first antenatal visit. *Journal of hypertension*. 2010 Jan;28(1):119-26. PubMed PMID: 19907344.
295. Green SM. When do clinical decision rules improve patient care? *Annals of emergency medicine*. 2013 Aug;62(2):132-5. PubMed PMID: 23548403.
296. Douma RA, Mos IC, Erkens PM, Nizet TA, Durian MF, Hovens MM, et al. Performance of 4 clinical decision rules in the diagnostic management of acute pulmonary embolism: a prospective cohort study. *Annals of internal medicine*. 2011 Jun 7;154(11):709-18. PubMed PMID: 21646554.
297. Than M, Flaws D, Sanders S, Doust J, Glasziou P, Kline J, et al. Development and validation of the Emergency Department Assessment of Chest pain Score and 2 h accelerated diagnostic protocol. *Emergency medicine Australasia : EMA*. 2014 Feb;26(1):34-44. PubMed PMID: 24428678.
298. Aldous SJ, Richards M, Cullen L, Troughton R, Than M. A 2-hour thrombolysis in myocardial infarction score outperforms other risk stratification tools in patients presenting with possible acute coronary syndromes: comparison of chest pain risk stratification tools. *American heart journal*. 2012 Oct;164(4):516-23. PubMed PMID: 23067909.
299. Than M, Herbert M, Flaws D, Cullen L, Hess E, Hollander JE, et al. What is an acceptable risk of major adverse cardiac event in chest pain patients soon after discharge from the Emergency Department?: a clinical survey. *International journal of cardiology*. 2013 Jul 1;166(3):752-4. PubMed PMID: 23084108.
300. Cowan N. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and brain sciences*. 2001 Feb;24(1):87-114; discussion -85. PubMed PMID: 11515286.
301. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*. 2008 Jan 30;27(2):157-72; discussion 207-12. PubMed PMID: 17569110.
302. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in medicine*. 2011 Jan 15;30(1):11-21. PubMed PMID: 21204120. Pubmed Central PMCID: 3341973.
303. Pickering JW, Endre ZH. New metrics for assessing diagnostic potential of candidate biomarkers. *Clinical journal of the American Society of Nephrology : CJASN*. 2012 Aug;7(8):1355-64. PubMed PMID: 22679181.
304. Than M, Cullen L, Aldous S, Parsonage WA, Reid CM, Greenslade J, et al. 2-Hour accelerated diagnostic protocol to assess patients with chest pain symptoms using contemporary troponins as the only biomarker: the ADAPT trial. *Journal of the American College of Cardiology*. 2012 Jun 5;59(23):2091-8. PubMed PMID: 22578923.
305. Bernstein SL, Aronsky D, Duseja R, Epstein S, Handel D, Hwang U, et al. The effect of emergency department crowding on clinically oriented outcomes. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*. 2009 Jan;16(1):1-10. PubMed PMID: 19007346.
306. Green SM, Schriger DL, Yealy DM. Methodologic Standards for Interpreting Clinical Decision Rules in Emergency Medicine: 2014 Update. *Annals of emergency medicine*. 2014 Sep;64(3):286-91. PubMed PMID: 24530108.
307. Dawes R, Corrigan B. Linear models in decision making. *Psychological Bulletin*. 1974;81(2):95-106.

308. Gibson NS, Sohne M, Kruip MJ, Tick LW, Gerdes VE, Bossuyt PM, et al. Further validation and simplification of the Wells clinical decision rule in pulmonary embolism. *Thrombosis and haemostasis*. 2008 Jan;99(1):229-34. PubMed PMID: 18217159.
309. Rothwell PM. Prognostic models. *Practical neurology*. 2008 Aug;8(4):242-53. PubMed PMID: 18644911.
310. Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P. Interpreting incremental value of markers added to risk prediction models. *American journal of epidemiology*. 2012 Sep 15;176(6):473-81. PubMed PMID: 22875755. Pubmed Central PMCID: 3530349.
311. Noble D, Dent T, Greenhalgh T. Time to compare impact and feasibility of prediction models in real life. *Bmj*. 2012;345:e4357; author reply e60. PubMed PMID: 22761093.
312. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology*. 2005 Oct;58(10):982-90. PubMed PMID: 16168343.
313. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS one*. 2008;3(8):e3081. PubMed PMID: 18769481. Pubmed Central PMCID: 2518111.
314. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *Jama*. 1999 Sep 15;282(11):1061-6. PubMed PMID: 10493205.
315. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2006 Feb 14;174(4):469-76. PubMed PMID: 16477057. Pubmed Central PMCID: 1373751.
316. Siersma V, Als-Nielsen B, Chen W, Hilden J, Gluud LL, Gluud C. Multivariable modelling for meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. *Statistics in medicine*. 2007 Jun 30;26(14):2745-58. PubMed PMID: 17117373.
317. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *Bmj*. 2012;344:e3318. PubMed PMID: 22628003.
318. Lugtenberg M, Zegers-van Schaick JM, Westert GP, Burgers JS. Why don't physicians adhere to guideline recommendations in practice? An analysis of barriers among Dutch general practitioners. *Implementation science : IS*. 2009;4:54. PubMed PMID: 19674440. Pubmed Central PMCID: 2734568.
319. Berner ES. Clinical decision support systems: state of the art. Rockville (MD): Agency for Healthcare Research and Quality; July 2009. AHRQ Publication No.:09-0054EF.
320. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Annals of internal medicine*. 2013 Apr 2;158(7):544-54. PubMed PMID: 23546566.
321. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working G. Systematic reviews of diagnostic test accuracy. *Annals of internal medicine*. 2008 Dec 16;149(12):889-97. PubMed PMID: 19075208. Pubmed Central PMCID: 2956514.
322. Ferrante di Ruffano L, Deeks JJ. Test-treatment RCTs are sheep in wolves' clothing. *Journal of clinical epidemiology*. 2015 Jun 28. PubMed PMID: 26130595.

323. Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. *Evidence-based medicine*. 2014 Apr;19(2):47-54. PubMed PMID: 24368333.
324. Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Systematic reviews*. 2012;1:60. PubMed PMID: 23194585. Pubmed Central PMCID: 3564748.
325. Perera R, Heneghan C, Yudkin P. Graphical method for depicting randomised trials of complex interventions. *Bmj*. 2007 Jan 20;334(7585):127-9. PubMed PMID: 17235093. Pubmed Central PMCID: 1779898.